

[A graph similarity for deep learning](#)

- Seongmin Ok
- abstract@[open-review](#): Graph neural networks (GNNs) have been successful in learning representations from graphs. Many popular GNNs follow the pattern of aggregate-transform: they aggregate the neighbors' attributes and then transform the results of aggregation with a learnable function. Analyses of these GNNs explain which pairs of non-identical graphs have different representations. However, we still lack an understanding of how similar these representations will be. We adopt kernel distance and propose transform-sum-cat as an alternative to aggregate-transform to reflect the continuous similarity between the node neighborhoods in the neighborhood aggregation. The idea leads to a simple and efficient graph similarity, which we name Weisfeiler-Leman similarity (WLS). In contrast to existing graph kernels, WLS is easy to implement with common deep learning frameworks. In graph classification experiments, transform-sum-cat significantly outperforms other neighborhood aggregation methods from popular GNN models. We also develop a simple and fast GNN model based on transform-sum-cat, which obtains, in comparison with widely used GNN models, (1) a higher accuracy in node classification, (2) a lower absolute error in graph regression, and (3) greater stability in adversarial training of graph generation.

[An Unsupervised Information-Theoretic Perceptual Quality Metric](#)

- Sangnie Bhardwaj, Ian Fischer, Johannes Ballā©, Troy Chinen
- abstract@[open-review](#): Tractable models of human perception have proved to be challenging to build. Hand-designed models such as MS-SSIM remain popular predictors of human image quality judgements due to their simplicity and speed. Recent modern deep learning approaches can perform better, but they rely on supervised data which can be costly to gather: large sets of class labels such as ImageNet, image quality ratings, or both. We combine recent advances in information-theoretic objective functions with a computational architecture informed by the physiology of the human visual system and unsupervised training on pairs of video frames, yielding our Perceptual Information Metric (PIM). We show that PIM is competitive with supervised metrics on the recent and challenging BAPPS image quality assessment dataset and outperforms them in predicting the ranking of image compression methods in CLIC 2020. We also perform qualitative experiments using the ImageNet-C dataset, and establish that PIM is robust with respect to architectural details.

[Self-Supervised MultiModal Versatile Networks](#)

- Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, Andrew Zisserman
- abstract@[open-review](#): Videos are a rich source of multi-modal supervision. In this work, we learn representations using self-supervision by leveraging three modalities naturally present in videos: visual, audio and language streams. To this end, we introduce the notion of a multimodal versatile network -- a network that can ingest multiple modalities and whose representations enable downstream tasks in multiple modalities. In particular, we explore how best to combine the modalities, such that fine-grained representations of the visual and audio modalities can be maintained, whilst also integrating text into a common embedding. Driven by versatility, we also introduce a novel process of deflation, so that the networks can be effortlessly applied to the visual data in the form of video or a static image. We demonstrate how such networks trained on large collections of unlabelled video data can be applied on video, video-text, image and audio tasks. Equipped with these representations, we obtain state-of-the-art performance on multiple challenging benchmarks including UCF101, HMDB51, Kinetics600, AudioSet and ESC-50 when compared to previous self-supervised work. Our models are publicly available.

[Benchmarking Deep Inverse Models over time, and the Neural-Adjoint method](#)

- Simiao Ren, Willie Padilla, Jordan Malof
- abstract@[open-review](#): We consider the task of solving generic inverse problems, where one wishes to determine the hidden parameters of a natural system that will give rise to a particular set of measurements. Recently many new approaches based upon deep learning have arisen, generating promising results. We conceptualize these models as different schemes for efficiently, but randomly, exploring the space of possible inverse solutions. As a result, the accuracy of each approach should be evaluated as a function of time rather than a single estimated solution, as is often done now. Using this metric, we compare several state-of-the-art inverse modeling approaches on four benchmark tasks: two existing tasks, a new 2-dimensional sinusoid task, and a challenging modern task of meta-material design. Finally, inspired by our conception of the inverse problem, we explore a simple solution that uses a deep neural network as a surrogate (i.e., approximation) for the forward model, and then uses backpropagation with respect to the model input to search for good inverse solutions. Variations of this approach - which we term the neural adjoint (NA) - have been explored recently on specific problems, and here we evaluate it comprehensively on our benchmark. We find that the addition of a simple novel loss term - which we term the boundary loss - dramatically improves the NA's performance, and it consequentially achieves the best (or nearly best) performance in all of our benchmark scenarios.

[Off-Policy Evaluation and Learning for External Validity under a Covariate Shift](#)

- Masatoshi Uehara, Masahiro Kato, Shota Yasui
- abstract@[open-review](#): We consider the evaluation and training of a new policy for the evaluation data by using the historical data obtained from a different policy. The goal of off-policy evaluation (OPE) is to estimate the expected reward of a new policy over the evaluation data, and that of off-policy learning (OPL) is to find a new policy that maximizes the expected reward over the evaluation data. Although the standard OPE and OPL assume the same distribution of covariate between the historical and evaluation data, there often exists a problem of a covariate shift, i.e., the distribution of the covariate of the historical data is different from that of the evaluation data. In this paper, we derive the efficiency bound of OPE under a covariate shift. Then, we propose doubly robust and efficient estimators for OPE and OPL under a covariate shift by using an estimator of the density ratio between the distributions of the historical and evaluation data. We also discuss other possible estimators and compare their theoretical properties. Finally, we confirm the effectiveness of the proposed estimators through experiments.

[Neural Methods for Point-wise Dependency Estimation](#)

- Yao-Hung Hubert Tsai, Han Zhao, Makoto Yamada, Louis-Philippe Morency, Russ R. Salakhutdinov
- abstract@[open-review](#): Since its inception, the neural estimation of mutual information (MI) has demonstrated the empirical success of modeling expected dependency between high-dimensional random variables. However, MI is an aggregate statistic and cannot be used to measure point-wise dependency between different events. In this work, instead of estimating the expected dependency, we focus on estimating point-wise dependency (PD), which quantitatively measures how likely two outcomes co-occur. We show that we can naturally obtain PD when we are optimizing MI neural variational bounds. However, optimizing these bounds is challenging due to its large variance in practice. To address this issue, we develop two methods (free of optimizing MI variational bounds): Probabilistic Classifier and Density-Ratio Fitting. We demonstrate the effectiveness of our approaches in 1) MI estimation, 2) self-supervised representation learning, and 3) cross-modal retrieval task.

[Fast and Flexible Temporal Point Processes with Triangular Maps](#)

- Oleksandr Shchur, Nicholas Gao, Marin Bilođ, Stephan Gähnemann
- abstract@[open-review](#): Temporal point process (TPP) models combined with recurrent neural networks provide a powerful framework for modeling continuous-time event data. While such models are flexible, they are inherently sequential and therefore cannot benefit from the parallelism of modern hardware. By exploiting the recent developments in the field of normalizing flows, we design TriTPP - a new class of non-recurrent TPP models, where

both sampling and likelihood computation can be done in parallel. TriTPP matches the flexibility of RNN-based methods but permits several orders of magnitude faster sampling. This enables us to use the new model for variational inference in continuous-time discrete-state systems. We demonstrate the advantages of the proposed framework on synthetic and real-world datasets.

[Backpropagating Linearly Improves Transferability of Adversarial Examples](#)

- Yiwen Guo, Qizhang Li, Hao Chen
- abstract@[open-review](#): The vulnerability of deep neural networks (DNNs) to adversarial examples has drawn great attention from the community. In this paper, we study the transferability of such examples, which lays the foundation of many black-box attacks on DNNs. We revisit a not so new but definitely noteworthy hypothesis of Goodfellow et al.'s and disclose that the transferability can be enhanced by improving the linearity of DNNs in an appropriate manner. We introduce linear backpropagation (LinBP), a method that performs backpropagation in a more linear fashion using off-the-shelf attacks that exploit gradients. More specifically, it calculates forward as normal but backpropagates loss as if some nonlinear activations are not encountered in the forward pass. Experimental results demonstrate that this simple yet effective method obviously outperforms current state-of-the-arts in crafting transferable adversarial examples on CIFAR-10 and ImageNet, leading to more effective attacks on a variety of DNNs. Code at: <https://github.com/qizhangli/linbp-attack>.

[PyGlove: Symbolic Programming for Automated Machine Learning](#)

- Daiyi Peng, Xuanyi Dong, Esteban Real, Mingxing Tan, Yifeng Lu, Gabriel Bender, Hanxiao Liu, Adam Kraft, Chen Liang, Quoc Le
- abstract@[open-review](#): Neural networks are sensitive to hyper-parameter and architecture choices. Automated Machine Learning (AutoML) is a promising paradigm for automating these choices. Current ML software libraries, however, are quite limited in handling the dynamic interactions among the components of AutoML. For example, efficient NAS algorithms, such as ENAS and DARTS, typically require an implementation coupling between the search space and search algorithm, the two key components in AutoML. Furthermore, implementing a complex search flow, such as searching architectures within a loop of searching hardware configurations, is difficult. To summarize, changing the search space, search algorithm, or search flow in current ML libraries usually requires a significant change in the program logic. In this paper, we introduce a new way of programming AutoML based on symbolic programming. Under this paradigm, ML programs are mutable, thus can be manipulated easily by another program. As a result, AutoML can be reformulated as an automated process of symbolic manipulation. With this formulation, we decouple the triangle of the search algorithm, the search space and the child program. This decoupling makes it easy to change the search space and search algorithm (without and with weight sharing), as well as to add search capabilities to existing code and implement complex search flows. We then introduce PyGlove, a new Python library that implements this paradigm. Through case studies on ImageNet and NAS-Bench-101, we show that with PyGlove users can easily convert a static program into a search space, quickly iterate on the search spaces and search algorithms, and craft complex search flows to achieve better results.

[Fourier Sparse Leverage Scores and Approximate Kernel Learning](#)

- Tamas Erdelyi, Cameron Musco, Christopher Musco
- abstract@[open-review](#): We prove new explicit upper bounds on the leverage scores of Fourier sparse functions under both the Gaussian and Laplace measures. In particular, we study s -sparse functions of the form $f(x) = \sum_{j=1}^s a_j e^{i\lambda_j x}$ for coefficients $a_j \in C$ and frequencies $\lambda_j \in R$. Bounding Fourier sparse leverage scores under various measures is of pure mathematical interest in approximation theory, and our work extends existing results for the uniform measure [Erd17, CP19a]. Practically, our bounds are motivated by two important applications in machine learning:
- Kernel Approximation. They yield a new random Fourier features algorithm for approximating Gaussian and Cauchy (rational quadratic) kernel matrices. For low-dimensional data, our method uses a near optimal number of features, and its runtime is polynomial in the *statistical dimension* of the approximated kernel matrix. It is the first ``oblivious sketching method'' with this property for any kernel besides the polynomial kernel, resolving an open question of [AKM+17, AKK+20b].
- Active Learning. They can be used as non-uniform sampling distributions for robust active learning when data follows a Gaussian or Laplace distribution. Using the framework of [AKM+19], we provide essentially optimal results for bandlimited and multiband interpolation, and Gaussian process regression. These results generalize existing work that only applies to uniformly distributed data.

[Improved Algorithms for Online Submodular Maximization via First-order Regret Bounds](#)

- Nicholas Harvey, Christopher Liaw, Tasuku Soma
- abstract@[open-review](#): We consider the problem of nonnegative submodular maximization in the online setting. At time step t , an algorithm selects a set $S_t \subseteq C$ where C is a feasible family of sets. An adversary then reveals a submodular function f_t . The goal is to design an efficient algorithm for minimizing the expected approximate regret. In this work, we give a general approach for improving regret bounds in online submodular maximization by exploiting first-order regret bounds for online linear optimization.
- For monotone submodular maximization subject to a matroid, we give an efficient algorithm which achieves a $(1 - c/e)^t \hat{\mu}$ -regret of $O(\hat{s}kT \ln(n/k))$ where n is the size of the ground set, k is the rank of the matroid, $\hat{\mu} > 0$ is a constant, and c is the average curvature. Even without assuming any curvature (i.e., taking $c = 1$), this regret bound improves on previous results of Streeter et al. (2009) and Golovin et al. (2014).
- For nonmonotone, unconstrained submodular functions, we give an algorithm with $1/2$ -regret $O(\hat{s}nT)$, improving on the results of Roughgarden and Wang (2018). Our approach is based on Blackwell approachability; in particular, we give a novel first-order regret bound for the Blackwell instances that arise in this setting

[Symbols: Probing Learning Algorithms with Synthetic Datasets](#)

- Alexandre Lacoste, Pau Rodríguez Lázquez, Frederic Branchaud-Charron, Parmida Atighchian, Massimo Caccia, Issam Hadj Laradji, Alexandre Drouin, Matthew Craddock, Laurent Charlin, David Vázquez
- abstract@[open-review](#): Progress in the field of machine learning has been fueled by the introduction of benchmark datasets pushing the limits of existing algorithms. Enabling the design of datasets to test specific properties and failure modes of learning algorithms is thus a problem of high interest, as it has a direct impact on innovation in the field. In this sense, we introduce Symbols – Synthetic Symbols – a tool for rapidly generating new datasets with a rich composition of latent features rendered in low resolution images. Symbols leverages the large amount of symbols available in the Unicode standard and the wide range of artistic font provided by the open font community. Our tool's high-level interface provides a language for rapidly generating new distributions on the latent features, including various types of textures and occlusions. To showcase the versatility of Symbols, we use it to dissect the limitations and flaws in standard learning algorithms in various learning setups including supervised learning, active learning, out of distribution generalization, unsupervised representation learning, and object counting.

[Adversarially Robust Streaming Algorithms via Differential Privacy](#)

- Avinatan Hassidim, Haim Kaplan, Yishay Mansour, Yossi Matias, Uri Stemmer

- abstract@[open-review](#): A streaming algorithm is said to be adversarially robust if its accuracy guarantees are maintained even when the data stream is chosen maliciously, by an adaptive adversary. We establish a connection between adversarial robustness of streaming algorithms and the notion of differential privacy. This connection allows us to design new adversarially robust streaming algorithms that outperform the current state-of-the-art constructions for many interesting regimes of parameters.

[Trading Personalization for Accuracy: Data Debugging in Collaborative Filtering](#)

- Long Chen, Yuan Yao, Feng Xu, Miao Xu, Hanghang Tong
- abstract@[open-review](#): Collaborative filtering has been widely used in recommender systems. Existing work has primarily focused on improving the prediction accuracy mainly via either building refined models or incorporating additional side information, yet has largely ignored the inherent distribution of the input rating data. In this paper, we propose a data debugging framework to identify overly personalized ratings whose existence degrades the performance of a given collaborative filtering model. The key idea of the proposed approach is to search for a small set of ratings whose editing (e.g., modification or deletion) would near-optimally improve the recommendation accuracy of a validation set. Experimental results demonstrate that the proposed approach can significantly improve the recommendation accuracy. Furthermore, we observe that the identified ratings significantly deviate from the average ratings of the corresponding items, and the proposed approach tends to modify them towards the average. This result sheds light on the design of future recommender systems in terms of balancing between the overall accuracy and personalization.

[Cascaded Text Generation with Markov Transformers](#)

- Yuntian Deng, Alexander Rush
- abstract@[open-review](#): The two dominant approaches to neural text generation are fully autoregressive models, using serial beam search decoding, and non-autoregressive models, using parallel decoding with no output dependencies. This work proposes an autoregressive model with sub-linear parallel time generation. Noting that conditional random fields with bounded context can be decoded in parallel, we propose an efficient cascaded decoding approach for generating high-quality output. To parameterize this cascade, we introduce a Markov transformer, a variant of the popular fully autoregressive model that allows us to simultaneously decode with specific autoregressive context cutoffs. This approach requires only a small modification from standard autoregressive training, while showing competitive accuracy/speed tradeoff compared to existing methods on five machine translation datasets.

[Improving Local Identifiability in Probabilistic Box Embeddings](#)

- Shib Dasgupta, Michael Boratko, Dongxu Zhang, Luke Vilnis, Xiang Li, Andrew McCallum
- abstract@[open-review](#): Geometric embeddings have recently received attention for their natural ability to represent transitive asymmetric relations via containment. Box embeddings, where objects are represented by n-dimensional hyperrectangles, are a particularly promising example of such an embedding as they are closed under intersection and their volume can be calculated easily, allowing them to naturally represent calibrated probability distributions. The benefits of geometric embeddings also introduce a problem of local identifiability, however, where whole neighborhoods of parameters result in equivalent loss which impedes learning. Prior work addressed some of these issues by using an approximation to Gaussian convolution over the box parameters, however this intersection operation also increases the sparsity of the gradient. In this work we model the box parameters with min and max Gumbel distributions, which were chosen such that the space is still closed under the operation of intersection. The calculation of the expected intersection volume involves all parameters, and we demonstrate experimentally that this drastically improves the ability of such models to learn.

[Permute-and-Flip: A new mechanism for differentially private selection](#)

- Ryan McKenna, Daniel R. Sheldon
- abstract@[open-review](#): We consider the problem of differentially private selection. Given a finite set of candidate items, and a quality score for each item, our goal is to design a differentially private mechanism that returns an item with a score that is as high as possible. The most commonly used mechanism for this task is the exponential mechanism. In this work, we propose a new mechanism for this task based on a careful analysis of the privacy constraints. The expected score of our mechanism is always at least as large as the exponential mechanism, and can offer improvements up to a factor of two. Our mechanism is simple to implement and runs in linear time.

[Deep reconstruction of strange attractors from time series](#)

- William Gilpin
- abstract@[open-review](#): Experimental measurements of physical systems often have a limited number of independent channels, causing essential dynamical variables to remain unobserved. However, many popular methods for unsupervised inference of latent dynamics from experimental data implicitly assume that the measurements have higher intrinsic dimensionality than the underlying system---making coordinate identification a dimensionality reduction problem. Here, we study the opposite limit, in which hidden governing coordinates must be inferred from only a low-dimensional time series of measurements. Inspired by classical analysis techniques for partial observations of chaotic attractors, we introduce a general embedding technique for univariate and multivariate time series, consisting of an autoencoder trained with a novel latent-space loss function. We show that our technique reconstructs the strange attractors of synthetic and real-world systems better than existing techniques, and that it creates consistent, predictive representations of even stochastic systems. We conclude by using our technique to discover dynamical attractors in diverse systems such as patient electrocardiograms, household electricity usage, neural spiking, and eruptions of the Old Faithful geyser---demonstrating diverse applications of our technique for exploratory data analysis.

[Reciprocal Adversarial Learning via Characteristic Functions](#)

- Shengxi Li, Zeyang Yu, Min Xiang, Danilo Mandic
- abstract@[open-review](#): Generative adversarial nets (GANs) have become a preferred tool for tasks involving complicated distributions. To stabilise the training and reduce the mode collapse of GANs, one of their main variants employs the integral probability metric (IPM) as the loss function. This provides extensive IPM-GANs with theoretical support for basically comparing moments in an embedded domain of the \textit{critic}. We generalise this by comparing the distributions rather than their moments via a powerful tool, i.e., the characteristic function (CF), which uniquely and universally comprising all the information about a distribution. For rigour, we first establish the physical meaning of the phase and amplitude in CF, and show that this provides a feasible way of balancing the accuracy and diversity of generation. We then develop an efficient sampling strategy to calculate the CFs. Within this framework, we further prove an equivalence between the embedded and data domains when a reciprocal exists, where we naturally develop the GAN in an auto-encoder structure, in a way of comparing everything in the embedded space (a semantically meaningful manifold). This efficient structure uses only two modules, together with a simple training strategy, to achieve bi-directionally generating clear images, which is referred to as the reciprocal CF GAN (RCF-GAN). Experimental results demonstrate the superior performances of the proposed RCF-GAN in terms of both generation and reconstruction.

[Statistical Guarantees of Distributed Nearest Neighbor Classification](#)

- Jixin Duan, Xingye Qiao, Guang Cheng

- abstract@[open-review](#): Nearest neighbor is a popular nonparametric method for classification and regression with many appealing properties. In the big data era, the sheer volume and spatial/temporal disparity of big data may prohibit centrally processing and storing the data. This has imposed considerable hurdle for nearest neighbor predictions since the entire training data must be memorized. One effective way to overcome this issue is the distributed learning framework. Through majority voting, the distributed nearest neighbor classifier achieves the same rate of convergence as its oracle version in terms of the regret, up to a multiplicative constant that depends solely on the data dimension. The multiplicative difference can be eliminated by replacing majority voting with the weighted voting scheme. In addition, we provide sharp theoretical upper bounds of the number of subsamples in order for the distributed nearest neighbor classifier to reach the optimal convergence rate. It is interesting to note that the weighted voting scheme allows a larger number of subsamples than the majority voting one. Our findings are supported by numerical studies.

Stein Self-Repulsive Dynamics: Benefits From Past Samples

- Mao Ye, Tongzheng Ren, Qiang Liu
- abstract@[open-review](#): We propose a new Stein self-repulsive dynamics for obtaining diversified samples from intractable un-normalized distributions. Our idea is to introduce Stein variational gradient as a repulsive force to push the samples of Langevin dynamics away from the past trajectories. This simple idea allows us to significantly decrease the auto-correlation in Langevin dynamics and hence increase the effective sample size. Importantly, as we establish in our theoretical analysis, the asymptotic stationary distribution remains correct even with the addition of the repulsive force, thanks to the special properties of the Stein variational gradient. We perform extensive empirical studies of our new algorithm, showing that our method yields much higher sample efficiency and better uncertainty estimation than vanilla Langevin dynamics.

The Statistical Complexity of Early-Stopped Mirror Descent

- Tomas Vaskevicius, Varun Kanade, Patrick Rebeschini
- abstract@[open-review](#): Recently there has been a surge of interest in understanding implicit regularization properties of iterative gradient-based optimization algorithms. In this paper, we study the statistical guarantees on the excess risk achieved by early-stopped unconstrained mirror descent algorithms applied to the unregularized empirical risk with the squared loss for linear models and kernel methods. By completing an inequality that characterizes convexity for the squared loss, we identify an intrinsic link between offset Rademacher complexities and potential-based convergence analysis of mirror descent methods. Our observation immediately yields excess risk guarantees for the path traced by the iterates of mirror descent in terms of offset complexities of certain function classes depending only on the choice of the mirror map, initialization point, step-size, and the number of iterations. We apply our theory to recover, in a rather clean and elegant manner via rather short proofs, some of the recent results in the implicit regularization literature, while also showing how to improve upon them in some settings.

Algorithmic recourse under imperfect causal knowledge: a probabilistic approach

- Amir-Hossein Karimi, Julius von Kägelgen, Bernhard Schölkopf, Isabel Valera
- abstract@[open-review](#): Recent work has discussed the limitations of counterfactual explanations to recommend actions for algorithmic recourse, and argued for the need of taking causal relationships between features into consideration. Unfortunately, in practice, the true underlying structural causal model is generally unknown. In this work, we first show that it is impossible to guarantee recourse without access to the true structural equations. To address this limitation, we propose two probabilistic approaches to select optimal actions that achieve recourse with high probability given limited causal knowledge (e.g., only the causal graph). The first captures uncertainty over structural equations under additive Gaussian noise, and uses Bayesian model averaging to estimate the counterfactual distribution. The second removes any assumptions on the structural equations by instead computing the average effect of recourse actions on individuals similar to the person who seeks recourse, leading to a novel subpopulation-based interventional notion of recourse. We then derive a gradient-based procedure for selecting optimal recourse actions, and empirically show that the proposed approaches lead to more reliable recommendations under imperfect causal knowledge than non-probabilistic baselines.

Quantitative Propagation of Chaos for SGD in Wide Neural Networks

- Valentin De Bortoli, Alain Durmus, Xavier Fontaine, Umut Simsekli
- abstract@[open-review](#): In this paper, we investigate the limiting behavior of a continuous-time counterpart of the Stochastic Gradient Descent (SGD) algorithm applied to two-layer overparameterized neural networks, as the number of neurons (i.e., the size of the hidden layer) $N \rightarrow \infty$. Following a probabilistic approach, we show 'propagation of chaos' for the particle system defined by this continuous-time dynamics under different scenarios, indicating that the statistical interaction between the particles asymptotically vanishes. In particular, we establish quantitative convergence with respect to N of any particle to a solution of a mean-field McKean-Vlasov equation in the metric space endowed with the Wasserstein distance. In comparison to previous works on the subject, we consider settings in which the sequence of stepsizes in SGD can potentially depend on the number of neurons and the iterations. We then identify two regimes under which different mean-field limits are obtained, one of them corresponding to an implicitly regularized version of the minimization problem at hand. We perform various experiments on real datasets to validate our theoretical results, assessing the existence of these two regimes on classification problems and illustrating our convergence results.

A Causal View on Robustness of Neural Networks

- Cheng Zhang, Kun Zhang, Yingzhen Li
- abstract@[open-review](#): We present a causal view on the robustness of neural networks against input manipulations, which applies not only to traditional classification tasks but also to general measurement data. Based on this view, we design a deep causal manipulation augmented model (deep CAMA) which explicitly models possible manipulations on certain causes leading to changes in the observed effect. We further develop data augmentation and test-time fine-tuning methods to improve deep CAMA's robustness. When compared with discriminative deep neural networks, our proposed model shows superior robustness against unseen manipulations. As a by-product, our model achieves disentangled representation which separates the representation of manipulations from those of other latent causes.

Minimax Classification with 0-1 Loss and Performance Guarantees

- Santiago Mazuelas, Andrea Zanoni, Aritz Pérez
- abstract@[open-review](#): Supervised classification techniques use training samples to find classification rules with small expected 0-1 loss. Conventional methods achieve efficient learning and out-of-sample generalization by minimizing surrogate losses over specific families of rules. This paper presents minimax risk classifiers (MRCs) that do not rely on a choice of surrogate loss and family of rules. MRCs achieve efficient learning and out-of-sample generalization by minimizing worst-case expected 0-1 loss w.r.t. uncertainty sets that are defined by linear constraints and include the true underlying distribution. In addition, MRCs' learning stage provides performance guarantees as lower and upper tight bounds for expected 0-1 loss. We also present MRCs' finite-sample generalization bounds in terms of training size and smallest minimax risk, and show their competitive classification performance w.r.t. state-of-the-art techniques using benchmark datasets.

How to Learn a Useful Critic? Model-based Action-Gradient-Estimator Policy Optimization

- Pierluca D'Oro, Wojciech Jastrząbowski

- abstract@[open-review](#): Deterministic-policy actor-critic algorithms for continuous control improve the actor by plugging its actions into the critic and ascending the action-value gradient, which is obtained by chaining the actor's Jacobian matrix with the gradient of the critic with respect to input actions. However, instead of gradients, the critic is, typically, only trained to accurately predict expected returns, which, on their own, are useless for policy optimization. In this paper, we propose MAGE, a model-based actor-critic algorithm, grounded in the theory of policy gradients, which explicitly learns the action-value gradient. MAGE backpropagates through the learned dynamics to compute gradient targets in temporal difference learning, leading to a critic tailored for policy improvement. On a set of MuJoCo continuous-control tasks, we demonstrate the efficiency of the algorithm in comparison to model-free and model-based state-of-the-art baselines.

[Coresets for Regressions with Panel Data](#)

- Lingxiao Huang, K Sudhir, Nisheeth Vishnoi
- abstract@[open-review](#): A panel dataset contains features or observations for multiple individuals over multiple time periods and regression problems with panel data are common in statistics and applied ML. When dealing with massive datasets, coresets have emerged as a valuable tool from a computational, storage and privacy perspective, as one needs to work with and share much smaller datasets. However, results on coresets for regression problems thus far have only been available for cross-sectional data (N individuals each observed for a single time unit) or longitudinal data (a single individual observed for $T > 1$ time units), but there are no results for panel data ($N > 1$, $T > 1$). This paper introduces the problem of coresets to panel data settings; we first define coresets for several variants of regression problems with panel data and then present efficient algorithms to construct coresets of size that are independent of N and T , and only polynomially depend on $1/\epsilon$ (where ϵ is the error parameter) and the number of regression parameters. Our approach is based on the Feldman-Langberg framework in which a key step is to upper bound the total sensitivity—that is roughly the sum of maximum influences of all individual-time pairs taken over all possible choices of regression parameters. Empirically, we assess our approach with a synthetic and a real-world datasets; the coreset sizes constructed using our approach are much smaller than the full dataset and coresets indeed accelerate the running time of computing the regression objective.

[Learning Composable Energy Surrogates for PDE Order Reduction](#)

- Alex Beatson, Jordan Ash, Geoffrey Roeder, Tianju Xue, Ryan P. Adams
- abstract@[open-review](#): Meta-materials are an important emerging class of engineered materials in which complex macroscopic behaviour—whether electromagnetic, thermal, or mechanical—arises from modular substructure. Simulation and optimization of these materials are computationally challenging, as rich substructures necessitate high-fidelity finite element meshes to solve the governing PDEs. To address this, we leverage parametric modular structure to learn component-level surrogates, enabling cheaper high-fidelity simulation. We use a neural network to model the stored potential energy in a component given boundary conditions. This yields a structured prediction task: macroscopic behavior is determined by the minimizer of the system's total potential energy, which can be approximated by composing these surrogate models. Composable energy surrogates thus permit simulation in the reduced basis of component boundaries. Costly ground-truth simulation of the full structure is avoided, as training data are generated by performing finite element analysis of individual components. Using dataset aggregation to choose training data allows us to learn energy surrogates which produce accurate macroscopic behavior when composed, accelerating simulation of parametric meta-materials.

[Efficient Contextual Bandits with Continuous Actions](#)

- Maryam Majzoubi, Chicheng Zhang, Rajan Chari, Akshay Krishnamurthy, John Langford, Aleksandrs Slivkins
- abstract@[open-review](#): We create a computationally tractable learning algorithm for contextual bandits with continuous actions having unknown structure. The new reduction-style algorithm composes with most supervised learning representations. We prove that this algorithm works in a general sense and verify the new functionality with large-scale experiments.

[Achieving Equalized Odds by Resampling Sensitive Attributes](#)

- Yaniv Romano, Stephen Bates, Emmanuel Candes
- abstract@[open-review](#): We present a flexible framework for learning predictive models that approximately satisfy the equalized odds notion of fairness. This is achieved by introducing a general discrepancy functional that rigorously quantifies violations of this criterion. This differentiable functional is used as a penalty driving the model parameters towards equalized odds. To rigorously evaluate fitted models, we develop a formal hypothesis test to detect whether a prediction rule violates this property, the first such test in the literature. Both the model fitting and hypothesis testing leverage a resampled version of the sensitive attribute obeying equalized odds, by construction. We demonstrate the applicability and validity of the proposed framework both in regression and multi-class classification problems, reporting improved performance over state-of-the-art methods. Lastly, we show how to incorporate techniques for equitable uncertainty quantification—unbiased for each group under study—to communicate the results of the data analysis in exact terms.

[Multi-Robot Collision Avoidance under Uncertainty with Probabilistic Safety Barrier Certificates](#)

- Wenhao Luo, Wen Sun, Ashish Kapoor
- abstract@[open-review](#): Safety in terms of collision avoidance for multi-robot systems is a difficult challenge under uncertainty, non-determinism, and lack of complete information. This paper aims to propose a collision avoidance method that accounts for both measurement uncertainty and motion uncertainty. In particular, we propose Probabilistic Safety Barrier Certificates (PrSBC) using Control Barrier Functions to define the space of admissible control actions that are probabilistically safe with formally provable theoretical guarantee. By formulating the chance constrained safety set into deterministic control constraints with PrSBC, the method entails minimally modifying an existing controller to determine an alternative safe controller via quadratic programming constrained to PrSBC constraints. The key advantage of the approach is that no assumptions about the form of uncertainty are required other than finite support, also enabling worst-case guarantees. We demonstrate effectiveness of the approach through experiments on realistic simulation environments.

[Hard Shape-Constrained Kernel Machines](#)

- Pierre-Cyril Aubin-Frankowski, Zoltan Szabo
- abstract@[open-review](#): Shape constraints (such as non-negativity, monotonicity, convexity) play a central role in a large number of applications, as they usually improve performance for small sample size and help interpretability. However enforcing these shape requirements in a hard fashion is an extremely challenging problem. Classically, this task is tackled (i) in a soft way (without out-of-sample guarantees), (ii) by specialized transformation of the variables on a case-by-case basis, or (iii) by using highly restricted function classes, such as polynomials or polynomial splines. In this paper, we prove that hard affine shape constraints on function derivatives can be encoded in kernel machines which represent one of the most flexible and powerful tools in machine learning and statistics. Particularly, we present a tightened second-order cone constrained reformulation, that can be readily implemented in convex solvers. We prove performance guarantees on the solution, and demonstrate the efficiency of the approach in joint quantile regression with applications to economics and to the analysis of aircraft trajectories, among others.

[A Closer Look at the Training Strategy for Modern Meta-Learning](#)

- JIAJIN CHEN, Xiao-Ming Wu, Yanke Li, Qimai LI, Li-Ming Zhan, Fu-lai Chung

- abstract@[open-review](#): The support/query (S/Q) episodic training strategy has been widely used in modern meta-learning algorithms and is believed to improve their generalization ability to test environments. This paper conducts a theoretical investigation of this training strategy on generalization. From a stability perspective, we analyze the generalization error bound of generic meta-learning algorithms trained with such strategy. We show that the S/Q episodic training strategy naturally leads to a counterintuitive generalization bound of $\mathcal{O}(1/\sqrt{n})$, which only depends on the task number n but independent of the inner-task sample size m . Under the common assumption $m <$

[On the Value of Out-of-Distribution Testing: An Example of Goodhart's Law](#)

- Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, Anton van den Hengel
- abstract@[open-review](#): Out-of-distribution (OOD) testing is increasingly popular for evaluating a machine learning system's ability to generalize beyond the biases of a training set. OOD benchmarks are designed to present a different joint distribution of data and labels between training and test time. VQA-CP has become the standard OOD benchmark for visual question answering, but we discovered three troubling practices in its current use. First, most published methods rely on explicit knowledge of the construction of the OOD splits. They often rely on inverting" the distribution of labels, e.g. answering "mostly yes" when the common training answer was ``no''. Second, the OOD test set is used for model selection. Third, a model's in-domain performance is assessed after retraining it on in-domain splits (VQA v2) that exhibit a more balanced distribution of labels. These three practices defeat the objective of evaluating generalization, and put into question the value of methods specifically designed for this dataset. We show that embarrassingly-simple methods, including one that generates answers at random, surpass the state of the art on some question types. We provide short- and long-term solutions to avoid these pitfalls and realize the benefits of OOD evaluation.

[Generalised Bayesian Filtering via Sequential Monte Carlo](#)

- Ayman Boustati, Omer Deniz Akyildiz, Theodoros Damoulas, Adam Johansen
- abstract@[open-review](#): We introduce a framework for inference in general state-space hidden Markov models (HMMs) under likelihood misspecification. In particular, we leverage the loss-theoretic perspective of Generalized Bayesian Inference (GBI) to define generalised filtering recursions in HMMs, that can tackle the problem of inference under model misspecification. In doing so, we arrive at principled procedures for robust inference against observation contamination by utilising the β -divergence. Operationalising the proposed framework is made possible via sequential Monte Carlo methods (SMC), where the standard particle methods, and their associated convergence results, are readily adapted to the new setting. We demonstrate our approach to object tracking and Gaussian process regression problems, and observe improved performance over standard filtering algorithms.

[Deterministic Approximation for Submodular Maximization over a Matroid in Nearly Linear Time](#)

- Kai Han, zongmai Cao, Shuang Cui, Benwei Wu
- abstract@[open-review](#): We study the problem of maximizing a non-monotone, non-negative submodular function subject to a matroid constraint. The prior best-known deterministic approximation ratio for this problem is $\frac{1}{4}\epsilon$ under $\mathcal{O}((n^4/\epsilon)\log n)$ time complexity. We show that this deterministic ratio can be improved to $\frac{1}{4}$ under $\mathcal{O}(nr)$ time complexity, and then present a more practical algorithm dubbed TwinGreedyFast which achieves $\frac{1}{4}\epsilon$ deterministic ratio in nearly-linear running time of $\mathcal{O}(\frac{n}{\epsilon}\log\frac{r}{\epsilon})$. Our approach is based on a novel algorithmic framework of simultaneously constructing two candidate solution sets through greedy search, which enables us to get improved performance bounds by fully exploiting the properties of independence systems. As a byproduct of this framework, we also show that TwinGreedyFast achieves $\frac{1}{2p+2}\epsilon$ deterministic ratio under a p -set system constraint with the same time complexity. To showcase the practicality of our approach, we empirically evaluated the performance of TwinGreedyFast on two network applications, and observed that it outperforms the state-of-the-art deterministic and randomized algorithms with efficient implementations for our problem.

[Flows for simultaneous manifold learning and density estimation](#)

- Johann Brehmer, Kyle Cranmer
- abstract@[open-review](#): We introduce manifold-learning flows ($\hat{\alpha}, \hat{\beta}$ -flows), a new class of generative models that simultaneously learn the data manifold as well as a tractable probability density on that manifold. Combining aspects of normalizing flows, GANs, autoencoders, and energy-based models, they have the potential to represent data sets with a manifold structure more faithfully and provide handles on dimensionality reduction, denoising, and out-of-distribution detection. We argue why such models should not be trained by maximum likelihood alone and present a new training algorithm that separates manifold and density updates. In a range of experiments we demonstrate how $\hat{\alpha}, \hat{\beta}$ -flows learn the data manifold and allow for better inference than standard flows in the ambient data space.

[Simultaneous Preference and Metric Learning from Paired Comparisons](#)

- Austin Xu, Mark Davenport
- abstract@[open-review](#): A popular model of preference in the context of recommendation systems is the so-called ideal point model. In this model, a user is represented as a vector u together with a collection of items $x_1 \dots x_N$ in a common low-dimensional space. The vector u represents the user's "ideal point," or the ideal combination of features that represents a hypothesized most preferred item. The underlying assumption in this model is that a smaller distance between u and an item x_j indicates a stronger preference for x_j . In the vast majority of the existing work on learning ideal point models, the underlying distance has been assumed to be Euclidean. However, this eliminates any possibility of interactions between features and a user's underlying preferences. In this paper, we consider the problem of learning an ideal point representation of a user's preferences when the distance metric is an unknown Mahalanobis metric. Specifically, we present a novel approach to estimate the user's ideal point u and the Mahalanobis metric from paired comparisons of the form "item x_i is preferred to item x_j ." This can be viewed as a special case of a more general metric learning problem where the location of some points are unknown a priori. We conduct extensive experiments on synthetic and real-world datasets to exhibit the effectiveness of our algorithm.

[Efficient Variational Inference for Sparse Deep Learning with Theoretical Guarantee](#)

- Jincheng Bai, Qifan Song, Guang Cheng
- abstract@[open-review](#): Sparse deep learning aims to address the challenge of huge storage consumption by deep neural networks, and to recover the sparse structure of target functions. Although tremendous empirical successes have been achieved, most sparse deep learning algorithms are lacking of theoretical supports. On the other hand, another line of works have proposed theoretical frameworks that are computationally infeasible. In this paper, we train sparse deep neural networks with a fully Bayesian treatment under spike-and-slab priors, and develop a set of computationally efficient variational inferences via continuous relaxation of Bernoulli distribution. The variational posterior contraction rate is provided, which justifies the consistency of the proposed variational Bayes method. Interestingly, our empirical results demonstrate that this variational procedure provides uncertainty quantification in terms of Bayesian predictive distribution and is also capable to accomplish consistent variable selection by training a sparse multi-layer neural network.

[Learning Manifold Implicitly via Explicit Heat-Kernel Learning](#)

- Yufan Zhou, Changyou Chen, Jinhui Xu

- abstract@[open-review](#): Manifold learning is a fundamental problem in machine learning with numerous applications. Most of the existing methods directly learn the low-dimensional embedding of the data in some high-dimensional space, and usually lack the flexibility of being directly applicable to down-stream applications. In this paper, we propose the concept of implicit manifold learning, where manifold information is implicitly obtained by learning the associated heat kernel. A heat kernel is the solution of the corresponding heat equation, which describes how ``heat'' transfers on the manifold, thus containing ample geometric information of the manifold. We provide both practical algorithm and theoretical analysis of our framework. The learned heat kernel can be applied to various kernel-based machine learning models, including deep generative models (DGM) for data generation and Stein Variational Gradient Descent for Bayesian inference. Extensive experiments show that our framework can achieve the state-of-the-art results compared to existing methods for the two tasks.

[Deep Relational Topic Modeling via Graph Poisson Gamma Belief Network](#)

- Chaojie Wang, Hao Zhang, Bo Chen, Dongsheng Wang, Zhengjue Wang, Mingyuan Zhou
- abstract@[open-review](#): To analyze a collection of interconnected documents, relational topic models (RTMs) have been developed to describe both the link structure and document content, exploring their underlying relationships via a single-layer latent representation with limited expressive capability. To better utilize the document network, we first propose graph Poisson factor analysis (GPFA) that constructs a probabilistic model for interconnected documents and also provides closed-form Gibbs sampling update equations, moving beyond sophisticated approximate assumptions of existing RTMs. Extending GPFA, we develop a novel hierarchical RTM named graph Poisson gamma belief network (GPGBN), and further introduce two different Weibull distribution based variational graph auto-encoders for efficient model inference and effective network information aggregation. Experimental results demonstrate that our models extract high-quality hierarchical latent document representations, leading to improved performance over baselines on various graph analytic tasks.

[One-bit Supervision for Image Classification](#)

- Hengtong Hu, Lingxi Xie, Zewei Du, Richang Hong, Qi Tian
- abstract@[open-review](#): This paper presents one-bit supervision, a novel setting of learning from incomplete annotations, in the scenario of image classification. Instead of training a model upon the accurate label of each sample, our setting requires the model to query with a predicted label of each sample and learn from the answer whether the guess is correct. This provides one bit (yes or no) of information, and more importantly, annotating each sample becomes much easier than finding the accurate label from many candidate classes. There are two keys to training a model upon one-bit supervision: improving the guess accuracy and making use of incorrect guesses. For these purposes, we propose a multi-stage training paradigm which incorporates negative label suppression into an off-the-shelf semi-supervised learning algorithm. In three popular image classification benchmarks, our approach claims higher efficiency in utilizing the limited amount of annotations.

[What is being transferred in transfer learning?](#)

- Behnam Neyshabur, Hanie Sedghi, Chiyuan Zhang
- abstract@[open-review](#): One desired capability for machines is the ability to transfer their understanding of one domain to another domain where data is (usually) scarce. Despite ample adaptation of transfer learning in many deep learning applications, we yet do not understand what enables a successful transfer and which part of the network is responsible for that. In this paper, we provide new tools and analysis to address these fundamental questions. Through a series of analysis on transferring to block-shuffled images, we separate the effect of feature reuse from learning high-level statistics of data and show that some benefit of transfer learning comes from the latter. We present that when training from pre-trained weights, the model stays in the same basin in the loss landscape and different instances of such model are similar in feature space and close in parameter space.

[Submodular Maximization Through Barrier Functions](#)

- Ashwinkumar Badanidiyuru, Amin Karbasi, Ehsan Kazemi, Jan Vondrak
- abstract@[open-review](#): In this paper, we introduce a novel technique for constrained submodular maximization, inspired by barrier functions in continuous optimization. This connection not only improves the running time for constrained submodular maximization but also provides the state of the art guarantee. More precisely, for maximizing a monotone submodular function subject to the combination of a $\$k\$$ -matchoid and $\$\ell\$$ -knapsack constraints (for $\$l \leq k\$$), we propose a potential function that can be approximately minimized. Once we minimize the potential function up to an $\$epsilon\$$ error, it is guaranteed that we have found a feasible set with a $\$2(k+1+\epsilon)\$$ -approximation factor which can indeed be further improved to $\$(k+1+\epsilon)\$$ by an enumeration technique. We extensively evaluate the performance of our proposed algorithm over several real-world applications, including a movie recommendation system, summarization tasks for YouTube videos, Twitter feeds and Yelp business locations, and a set cover problem.

[Neural Networks with Recurrent Generative Feedback](#)

- Yujia Huang, James Gornet, Sihui Dai, Zhiding Yu, Tan Nguyen, Doris Tsao, Anima Anandkumar
- abstract@[open-review](#): Neural networks are vulnerable to input perturbations such as additive noise and adversarial attacks. In contrast, human perception is much more robust to such perturbations. The Bayesian brain hypothesis states that human brains use an internal generative model to update the posterior beliefs of the sensory input. This mechanism can be interpreted as a form of self-consistency between the maximum a posteriori (MAP) estimation of an internal generative model and the external environment. Inspired by such hypothesis, we enforce self-consistency in neural networks by incorporating generative recurrent feedback. We instantiate this design on convolutional neural networks (CNNs). The proposed framework, termed Convolutional Neural Networks with Feedback (CNN-F), introduces a generative feedback with latent variables to existing CNN architectures, where consistent predictions are made through alternating MAP inference under a Bayesian framework. In the experiments, CNN-F shows considerably improved adversarial robustness over conventional feedforward CNNs on standard benchmarks.

[Learning to Extrapolate Knowledge: Transductive Few-shot Out-of-Graph Link Prediction](#)

- Jinheon Baek, Dong Bok Lee, Sung Ju Hwang
- abstract@[open-review](#): Many practical graph problems, such as knowledge graph construction and drug-drug interaction prediction, require to handle multi-relational graphs. However, handling real-world multi-relational graphs with Graph Neural Networks (GNNs) is often challenging due to their evolving nature, as new entities (nodes) can emerge over time. Moreover, newly emerged entities often have few links, which makes the learning even more difficult. Motivated by this challenge, we introduce a realistic problem of few-shot out-of-graph link prediction, where we not only predict the links between the seen and unseen nodes as in a conventional out-of-knowledge link prediction task but also between the unseen nodes, with only few edges per node. We tackle this problem with a novel transductive meta-learning framework which we refer to as Graph Extrapolation Networks (GEN). GEN meta-learns both the node embedding network for inductive inference (seen-to-unseen) and the link prediction network for transductive inference (unseen-to-unseen). For transductive link prediction, we further propose a stochastic embedding layer to model uncertainty in the link prediction between unseen entities. We validate our model on multiple benchmark datasets for knowledge graph completion and drug-drug interaction prediction. The results show that our model significantly outperforms relevant baselines for out-of-graph link prediction tasks.

[Exploiting weakly supervised visual patterns to learn from partial annotations](#)

- Kaustav Kundu, Joseph Tighe
- abstract@[open-review](#): As classifications datasets progressively get larger in terms of label space and number of examples, annotating them with all labels becomes non-trivial and expensive task. For example, annotating the entire OpenImage test set can cost \$6.5M. Hence, in current large-scale benchmarks such as OpenImages and LVIS, less than 1% of the labels are annotated across all images. Standard classification models are trained in a manner where these un-annotated labels are ignored. Ignoring these un-annotated labels result in loss of supervisory signal which reduces the performance of the classification models. Instead, in this paper, we exploit relationships among images and labels to derive more supervisory signal from the un-annotated labels. We study the effectiveness of our approach across several multi-label computer vision benchmarks, such as CIFAR100, MS-COCO panoptic segmentation, OpenImage and LVIS datasets. Our approach can outperform baselines by a margin of 2-10% across all the datasets on mean average precision (mAP) and mean F1 metrics.

[Improving Inference for Neural Image Compression](#)

- Yibo Yang, Robert Bamler, Stephan Mandt
- abstract@[open-review](#): We consider the problem of lossy image compression with deep latent variable models. State-of-the-art methods build on hierarchical variational autoencoders (VAEs) and learn inference networks to predict a compressible latent representation of each data point. Drawing on the variational inference perspective on compression, we identify three approximation gaps which limit performance in the conventional approach: an amortization gap, a discretization gap, and a marginalization gap. We propose remedies for each of these three limitations based on ideas related to iterative inference, stochastic annealing for discrete optimization, and bits-back coding, resulting in the first application of bits-back coding to lossy compression. In our experiments, which include extensive baseline comparisons and ablation studies, we achieve new state-of-the-art performance on lossy image compression using an established VAE architecture, by changing only the inference method.

[Neuron Merging: Compensating for Pruned Neurons](#)

- Woojeong Kim, Suhyun Kim, Mincheol Park, Geunseok Jeon
- abstract@[open-review](#): Network pruning is widely used to lighten and accelerate neural network models. Structured network pruning discards the whole neuron or filter, leading to accuracy loss. In this work, we propose a novel concept of neuron merging applicable to both fully connected layers and convolution layers, which compensates for the information loss due to the pruned neurons/filters. Neuron merging starts with decomposing the original weights into two matrices/tensors. One of them becomes the new weights for the current layer, and the other is what we name a scaling matrix, guiding the combination of neurons. If the activation function is ReLU, the scaling matrix can be absorbed into the next layer under certain conditions, compensating for the removed neurons. We also propose a data-free and inexpensive method to decompose the weights by utilizing the cosine similarity between neurons. Compared to the pruned model with the same topology, our merged model better preserves the output feature map of the original model; thus, it maintains the accuracy after pruning without fine-tuning. We demonstrate the effectiveness of our approach over network pruning for various model architectures and datasets. As an example, for VGG-16 on CIFAR-10, we achieve an accuracy of 93.16% while reducing 64% of total parameters, without any fine-tuning. The code can be found here: <https://github.com/friendshipkim/neuron-merging>

[FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence](#)

- Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A. Raffel, Ekin Dogus Cubuk, Alexey Kurakin, Chun-Liang Li
- abstract@[open-review](#): Semi-supervised learning (SSL) provides an effective means of leveraging unlabeled data to improve a model's performance. This domain has seen fast progress recently, at the cost of requiring more complex methods. In this paper we propose FixMatch, an algorithm that is a significant simplification of existing SSL methods. FixMatch first generates pseudo-labels using the model's predictions on weakly-augmented unlabeled images. For a given image, the pseudo-label is only retained if the model produces a high-confidence prediction. The model is then trained to predict the pseudo-label when fed a strongly-augmented version of the same image. Despite its simplicity, we show that FixMatch achieves state-of-the-art performance across a variety of standard semi-supervised learning benchmarks, including 94.93% accuracy on CIFAR-10 with 250 labels and 88.61% accuracy with 40 just 4 labels per class. We carry out an extensive ablation study to tease apart the experimental factors that are most important to FixMatch's success. The code is available at <https://github.com/google-research/fixmatch>.

[Reinforcement Learning with Combinatorial Actions: An Application to Vehicle Routing](#)

- Arthur Delarue, Ross Anderson, Christian Tjandraatmadja
- abstract@[open-review](#): Value-function-based methods have long played an important role in reinforcement learning. However, finding the best next action given a value function of arbitrary complexity is nontrivial when the action space is too large for enumeration. We develop a framework for value-function-based deep reinforcement learning with a combinatorial action space, in which the action selection problem is explicitly formulated as a mixed-integer optimization problem. As a motivating example, we present an application of this framework to the capacitated vehicle routing problem (CVRP), a combinatorial optimization problem in which a set of locations must be covered by a single vehicle with limited capacity. On each instance, we model an action as the construction of a single route, and consider a deterministic policy which is improved through a simple policy iteration algorithm. Our approach is competitive with other reinforcement learning methods and achieves an average gap of 1.7% with state-of-the-art OR methods on standard library instances of medium size.

[Towards Playing Full MOBA Games with Deep Reinforcement Learning](#)

- Deheng Ye, Guibin Chen, Wen Zhang, Sheng Chen, Bo Yuan, Bo Liu, Jia Chen, Zhao Liu, Fuhao Qiu, Hongsheng Yu, Yinyuting Yin, Bei Shi, Liang Wang, Tengfei Shi, Qiang Fu, Wei Yang, Lanxiao Huang, Wei Liu
- abstract@[open-review](#): MOBA games, e.g., Honor of Kings, League of Legends, and Dota 2, pose grand challenges to AI systems such as multi-agent, enormous state-action space, complex action control, etc. Developing AI for playing MOBA games has raised much attention accordingly. However, existing work falls short in handling the raw game complexity caused by the explosion of agent combinations, i.e., lineups, when expanding the hero pool in case that OpenAI's Dota AI limits the play to a pool of only 17 heroes. As a result, full MOBA games without restrictions are far from being mastered by any existing AI system. In this paper, we propose a MOBA AI learning paradigm that methodologically enables playing full MOBA games with deep reinforcement learning. Specifically, we develop a combination of novel and existing learning techniques, including off-policy adaption, multi-head value estimation, curriculum self-play learning, policy distillation, and Monte-Carlo tree-search, in training and playing a large pool of heroes, meanwhile addressing the scalability issue skillfully. Tested on Honor of Kings, a popular MOBA game, we show how to build superhuman AI agents that can defeat top esports players. The superiority of our AI is demonstrated by the first large-scale performance test of MOBA AI agent in the literature.

[Rankmax: An Adaptive Projection Alternative to the Softmax Function](#)

- Weiwei Kong, Walid Krichene, Nicolas Mayoraz, Steffen Rendle, Li Zhang
- abstract@[open-review](#): Several machine learning models involve mapping a score vector to a probability vector. Usually, this is done by projecting the score vector onto a probability simplex, and such projections are often characterized as Lipschitz continuous approximations of the argmax function, whose Lipschitz constant is controlled by a parameter that is similar to a softmax temperature. The aforementioned parameter has been observed to affect the quality of these models and is typically either treated as a constant or decayed over time. In this work, we propose a method that adapts this parameter to individual training examples. The resulting method exhibits desirable properties, such as sparsity of its support and numerically efficient

implementation, and we find that it significantly outperforms competing non-adaptive projection methods. In our analysis, we also derive the general solution of (Bregman) projections onto the (n, k) -simplex, a result which may be of independent interest.

[Online Agnostic Boosting via Regret Minimization](#)

- Nataly Brukhim, Xinyi Chen, Elad Hazan, Shay Moran
- abstract@[open-review](#): Boosting is a widely used machine learning approach based on the idea of aggregating weak learning rules. While in statistical learning numerous boosting methods exist both in the realizable and agnostic settings, in online learning they exist only in the realizable case. In this work we provide the first agnostic online boosting algorithm; that is, given a weak learner with only marginally-better-than-trivial regret guarantees, our algorithm boosts it to a strong learner with sublinear regret. Our algorithm is based on an abstract (and simple) reduction to online convex optimization, which efficiently converts an arbitrary online convex optimizer to an online booster. Moreover, this reduction extends to the statistical as well as the online realizable settings, thus unifying the 4 cases of statistical/online and agnostic/realizable boosting.

[Causal Intervention for Weakly-Supervised Semantic Segmentation](#)

- Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, Qianru Sun
- abstract@[open-review](#): We present a causal inference framework to improve Weakly-Supervised Semantic Segmentation (WSSS). Specifically, we aim to generate better pixel-level pseudo-masks by using only image-level labels -- the most crucial step in WSSS. We attribute the cause of the ambiguous boundaries of pseudo-masks to the confounding context, e.g., the correct image-level classification of "horse" and "person" may be not only due to the recognition of each instance, but also their co-occurrence context, making the model inspection (e.g., CAM) hard to distinguish between the boundaries. Inspired by this, we propose a structural causal model to analyze the causalities among images, contexts, and class labels. Based on it, we develop a new method: Context Adjustment (CONTA), to remove the confounding bias in image-level classification and thus provide better pseudo-masks as ground-truth for the subsequent segmentation model. On PASCAL VOC 2012 and MS-COCO, we show that CONTA boosts various popular WSSS methods to new state-of-the-arts.

[Belief Propagation Neural Networks](#)

- Jonathan Kuck, Shuvam Chakraborty, Hao Tang, Rachel Luo, Jiaming Song, Ashish Sabharwal, Stefano Ermon
- abstract@[open-review](#): Learned neural solvers have successfully been used to solve combinatorial optimization and decision problems. More general counting variants of these problems, however, are still largely solved with hand-crafted solvers. To bridge this gap, we introduce belief propagation neural networks (BPNNs), a class of parameterized operators that operate on factor graphs and generalize Belief Propagation (BP). In its strictest form, a BPNN layer (BPNN-D) is a learned iterative operator that provably maintains many of the desirable properties of BP for any choice of the parameters. Empirically, we show that by training BPNN-D learns to perform the task better than the original BP: it converges 1.7x faster on Ising models while providing tighter bounds. On challenging model counting problems, BPNNs compute estimates 100's of times faster than state-of-the-art handcrafted methods, while returning an estimate of comparable quality.

[Over-parameterized Adversarial Training: An Analysis Overcoming the Curse of Dimensionality](#)

- Yi Zhang, Orestis Plevrakis, Simon S. Du, Xingguo Li, Zhao Song, Sanjeev Arora
- abstract@[open-review](#): Adversarial training is a popular method to give neural nets robustness against adversarial perturbations. In practice adversarial training leads to low robust training loss. However, a rigorous explanation for why this happens under natural conditions is still missing. Recently a convergence theory of standard (non-adversarial) supervised training was developed by various groups for {\em very overparametrized} nets. It is unclear how to extend these results to adversarial training because of the min-max objective. Recently, a first step towards this direction was made by Gao et al. using tools from online learning, but they require the width of the net to be \emph{exponential} in input dimension \$d\$, and with an unnatural activation function. Our work proves convergence to low robust training loss for \emph{polynomial} width instead of exponential, under natural assumptions and with ReLU activations. A key element of our proof is showing that ReLU networks near initialization can approximate the step function, which may be of independent interest.

[Post-training Iterative Hierarchical Data Augmentation for Deep Networks](#)

- Adil Khan, Khadija Fraz
- abstract@[open-review](#): In this paper, we propose a new iterative hierarchical data augmentation (IHDA) method to fine-tune trained deep neural networks to improve their generalization performance. The IHDA is motivated by three key insights: (1) Deep networks (DNs) are good at learning multi-level representations from data. (2) Performing data augmentation (DA) in the learned feature spaces of DNs can significantly improve their performance. (3) Implementing DA in hard-to-learn regions of a feature space can effectively augment the dataset to improve generalization. Accordingly, the IHDA performs DA in a deep feature space, at level 1, by transforming it into a distribution space and synthesizing new samples using the learned distributions for data points that lie in hard-to-classify regions, which is estimated by analyzing the neighborhood characteristics of each data point. The synthesized samples are used to fine-tune the parameters of the subsequent layers. The same procedure is then repeated for the feature space at level l+1. To avoid overfitting, the concept of dropout probability is employed, which is gradually relaxed as the IHDA works towards high-level feature spaces. IHDA provided a state-of-the-art performance on CIFAR-10, CIFAR-100, and ImageNet for several DNs, and beat the performance of existing state-of-the-art DA approaches for the same networks on these datasets. Finally, to demonstrate its domain-agnostic properties, we show the significant improvements that IHDA provided for a deep neural network on a non-image wearable sensor-based activity recognition benchmark.

[Debugging Tests for Model Explanations](#)

- Julius Adebayo, Michael Muelly, Ilaria Liccardi, Been Kim
- abstract@[open-review](#): We investigate whether post-hoc model explanations are effective for diagnosing model errors--model debugging. In response to the challenge of explaining a model's prediction, a vast array of explanation methods have been proposed. Despite increasing use, it is unclear if they are effective. To start, we categorize \textit{bugs}, based on their source, into: \textit{data, model, and test-time} contamination bugs. For several explanation methods, we assess their ability to: detect spurious correlation artifacts (data contamination), diagnose mislabeled training examples (data contamination), differentiate between a (partially) re-initialized model and a trained one (model contamination), and detect out-of-distribution inputs (test-time contamination). We find that the methods tested are able to diagnose a spurious background bug, but not conclusively identify mislabeled training examples. In addition, a class of methods, that modify the back-propagation algorithm are invariant to the higher layer parameters of a deep network; hence, ineffective for diagnosing model contamination. We complement our analysis with a human subject study, and find that subjects fail to identify defective models using attributions, but instead rely, primarily, on model predictions. Taken together, our results provide guidance for practitioners and researchers turning to explanations as tools for model debugging.

[Robust compressed sensing using generative models](#)

- Ajil Jalal, Liu Liu, Alexandros G. Dimakis, Constantine Caramanis

- abstract@[open-review](#): We consider estimating a high dimensional signal in \mathbb{R}^n using a sublinear number of linear measurements. In analogy to classical compressed sensing, here we assume a generative model as a prior, that is, we assume the signal is represented by a deep generative model $G: \mathbb{R}^k \rightarrow \mathbb{R}^n$. Classical recovery approaches such as empirical risk minimization (ERM) are guaranteed to succeed when the measurement matrix is sub-Gaussian. However, when the measurement matrix and measurements are heavy tailed or have outliers, recovery may fail dramatically. In this paper we propose an algorithm inspired by the Median-of-Means (MOM). Our algorithm guarantees recovery for heavy tailed data, even in the presence of outliers. Theoretically, our results show our novel MOM-based algorithm enjoys the same sample complexity guarantees as ERM under sub-Gaussian assumptions. Our experiments validate both aspects of our claims: other algorithms are indeed fragile and fail under heavy tailed and/or corrupted data, while our approach exhibits the predicted robustness.

Fairness without Demographics through Adversarially Reweighted Learning

- Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, Ed Chi
- abstract@[open-review](#): Much of the previous machine learning (ML) fairness literature assumes that protected features such as race and sex are present in the dataset, and relies upon them to mitigate fairness concerns. However, in practice factors like privacy and regulation often preclude the collection of protected features, or their use for training or inference, severely limiting the applicability of traditional fairness research. Therefore, we ask: How can we train a ML model to improve fairness when we do not even know the protected group memberships? In this work we address this problem by proposing Adversarially Reweighted Learning (ARL). In particular, we hypothesize that non-protected features and task labels are valuable for identifying fairness issues, and can be used to co-train an adversarial reweighting approach for improving fairness. Our results show that ARL improves Rawlsian Max-Min fairness, with notable AUC improvements for worst-case protected groups in multiple datasets, outperforming state-of-the-art alternatives.

Stochastic Latent Actor-Critic: Deep Reinforcement Learning with a Latent Variable Model

- Alex X. Lee, Anusha Nagabandi, Pieter Abbeel, Sergey Levine
- abstract@[open-review](#): Deep reinforcement learning (RL) algorithms can use high-capacity deep networks to learn directly from image observations. However, these high-dimensional observation spaces present a number of challenges in practice, since the policy must now solve two problems: representation learning and task learning. In this work, we tackle these two problems separately, by explicitly learning latent representations that can accelerate reinforcement learning from images. We propose the stochastic latent actor-critic (SLAC) algorithm: a sample-efficient and high-performing RL algorithm for learning policies for complex continuous control tasks directly from high-dimensional image inputs. SLAC provides a novel and principled approach for unifying stochastic sequential models and RL into a single method, by learning a compact latent representation and then performing RL in the model's learned latent space. Our experimental evaluation demonstrates that our method outperforms both model-free and model-based alternatives in terms of final performance and sample efficiency, on a range of difficult image-based control tasks. Our code and videos of our results are available at our website.

Ridge Rider: Finding Diverse Solutions by Following Eigenvectors of the Hessian

- Jack Parker-Holder, Luke Metz, Cinjon Resnick, Hengyuan Hu, Adam Lerer, Alistair Letcher, Alexander Peysakhovich, Aldo Pacchiano, Jakob Foerster
- abstract@[open-review](#): Over the last decade, a single algorithm has changed many facets of our lives - Stochastic Gradient Descent (SGD). In the era of ever decreasing loss functions, SGD and its various offspring have become the go-to optimization tool in machine learning and are a key component of the success of deep neural networks (DNNs). While SGD is guaranteed to converge to a local optimum (under loose assumptions), in some cases it may matter which local optimum is found, and this is often context-dependent. Examples frequently arise in machine learning, from shape-versus-texture-features to ensemble methods and zero-shot coordination. In these settings, there are desired solutions which SGD on standard' loss functions will not find, since it instead converges to the easy' solutions. In this paper, we present a different approach. Rather than following the gradient, which corresponds to a locally greedy direction, we instead follow the eigenvectors of the Hessian. By iteratively following and branching amongst the ridges, we effectively span the loss surface to find qualitatively different solutions. We show both theoretically and experimentally that our method, called Ridge Rider (RR), offers a promising direction for a variety of challenging problems.

The route to chaos in routing games: When is price of anarchy too optimistic?

- Thiparat Chotibut, Fryderyk Falniowski, MichaÅ, Misiurewicz, Georgios Piliouras
- abstract@[open-review](#): Routing games are amongst the most studied classes of games in game theory. Their most well-known property is that learning dynamics typically converge to equilibria implying approximately optimal performance (low Price of Anarchy). We perform a stress test for these classic results by studying the ubiquitous learning dynamics, Multiplicative Weights Update (MWU), in different classes of congestion games, uncovering intricate non-equilibrium phenomena. We study MWU using the actual game costs without applying cost normalization to $[0,1]$. Although this non-standard assumption leads to large regret, it captures realistic agents' behaviors. Namely, as the total demand increases, agents respond more aggressively to unbearably large costs.

We start with the illustrative case of non-atomic routing games with two paths of linear cost, and show that every system has a carrying capacity, above which it becomes unstable. If the equilibrium flow is a symmetric 50-50% split, the system exhibits one period-doubling bifurcation. Although the Price of Anarchy is equal to one, in the large population limit the time-average social cost for all but a zero measure set of initial conditions converges to its worst possible value. For asymmetric equilibrium flows, increasing the demand eventually forces the system into Li-Yorke chaos with positive topological entropy and periodic orbits of all possible periods. Remarkably, in all non-equilibrating regimes, the time-average flows on the paths converge exactly to the equilibrium flows, a property akin to no-regret learning in zero-sum games. We extend our results to games with arbitrarily many strategies, polynomial cost functions, non-atomic as well as atomic routing games, and heterogenous users.

Online Algorithm for Unsupervised Sequential Selection with Contextual Information

- Arun Verma, Manjesh Kumar Hanawal, Csaba Szepesvari, Venkatesh Saligrama
- abstract@[open-review](#): In this paper, we study Contextual Unsupervised Sequential Selection (USS), a new variant of the stochastic contextual bandits problem where the loss of an arm cannot be inferred from the observed feedback. In our setup, arms are associated with fixed costs and are ordered, forming a cascade. In each round, a context is presented, and the learner selects the arms sequentially till some depth. The total cost incurred by stopping at an arm is the sum of fixed costs of arms selected and the stochastic loss associated with the arm. The learner's goal is to learn a decision rule that maps contexts to arms with the goal of minimizing the total expected loss. The problem is challenging as we are faced with an unsupervised setting as the total loss cannot be estimated. Clearly, learning is feasible only if the optimal arm can be inferred (explicitly or implicitly) from the problem structure. We observe that learning is still possible when the problem instance satisfies the so-called 'Contextual Weak Dominance' (CWD) property. Under CWD, we propose an algorithm for the contextual USS problem and demonstrate that it has sub-linear regret. Experiments on synthetic and real datasets validate our algorithm.

Adapting Neural Architectures Between Domains

- Yanxi Li, Zhaohui Yang, Yunhe Wang, Chang Xu
- abstract@[open-review](#): Neural architecture search (NAS) has demonstrated impressive performance in automatically designing high-performance neural networks. The power of deep neural networks is to be unleashed for analyzing a large volume of data (e.g. ImageNet), but the architecture search is often

executed on another smaller dataset (e.g. CIFAR-10) to finish it in a feasible time. However, it is hard to guarantee that the optimal architecture derived on the proxy task could maintain its advantages on another more challenging dataset. This paper aims to improve the generalization of neural architectures via domain adaptation. We analyze the generalization bounds of the derived architecture and suggest its close relations with the validation error and the data distribution distance on both domains. These theoretical analyses lead to AdaptNAS, a novel and principled approach to adapt neural architectures between domains in NAS. Our experimental evaluation shows that only a small part of ImageNet will be sufficient for AdaptNAS to extend its architecture success to the entire ImageNet and outperform state-of-the-art comparison algorithms.

[What went wrong and when? Instance-wise feature importance for time-series black-box models](#)

- Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David K. Duvenaud, Anna Goldenberg
- abstract@[open-review](#): Explanations of time series models are useful for high stakes applications like healthcare but have received little attention in machine learning literature. We propose FIT, a framework that evaluates the importance of observations for a multivariate time-series black-box model by quantifying the shift in the predictive distribution over time. FIT defines the importance of an observation based on its contribution to the distributional shift under a KL-divergence that contrasts the predictive distribution against a counterfactual where the rest of the features are unobserved. We also demonstrate the need to control for time-dependent distribution shifts. We compare with state-of-the-art baselines on simulated and real-world clinical data and demonstrate that our approach is superior in identifying important time points and observations throughout the time series.

[Towards Better Generalization of Adaptive Gradient Methods](#)

- Yingxue Zhou, Belhal Karimi, Jinxing Yu, Zhiqiang Xu, Ping Li
- abstract@[open-review](#): Adaptive gradient methods such as AdaGrad, RMSprop and Adam have been optimizers of choice for deep learning due to their fast training speed. However, it was recently observed that their generalization performance is often worse than that of SGD for over-parameterized neural networks. While new algorithms such as AdaBound, SWAT, and Padam were proposed to improve the situation, the provided analyses are only committed to optimization bounds for the training objective, leaving critical generalization capacity unexplored. To close this gap, we propose \textit{adaptive gradient} for nonconvex optimization which leverages differential privacy to boost the generalization performance of adaptive gradient methods. Theoretical analyses show that \textit{SAGD} has high-probability convergence to a population stationary point. We further conduct experiments on various popular deep learning tasks and models. Experimental results illustrate that \textit{SAGD} is empirically competitive and often better than baselines.

[Learning Guidance Rewards with Trajectory-space Smoothing](#)

- Tanmay Gangwani, Yuan Zhou, Jian Peng
- abstract@[open-review](#): Long-term temporal credit assignment is an important challenge in deep reinforcement learning (RL). It refers to the ability of the agent to attribute actions to consequences that may occur after a long time interval. Existing policy-gradient and Q-learning algorithms typically rely on dense environmental rewards that provide rich short-term supervision and help with credit assignment. However, they struggle to solve tasks with delays between an action and the corresponding rewarding feedback. To make credit assignment easier, recent works have proposed algorithms to learn dense "guidance" rewards that could be used in place of the sparse or delayed environmental rewards. This paper is in the same vein -- starting with a surrogate RL objective that involves smoothing in the trajectory-space, we arrive at a new algorithm for learning guidance rewards. We show that the guidance rewards have an intuitive interpretation, and can be obtained without training any additional neural networks. Due to the ease of integration, we use the guidance rewards in a few popular algorithms (Q-learning, Actor-Critic, Distributional-RL) and present results in single-agent and multi-agent tasks that elucidate the benefit of our approach when the environmental rewards are sparse or delayed.

[Variance Reduction via Accelerated Dual Averaging for Finite-Sum Optimization](#)

- Chaobing Song, Yong Jiang, Yi Ma
- abstract@[open-review](#): In this paper, we introduce a simplified and unified method for finite-sum convex optimization, named \textbf{VRADA}. In the general convex and smooth setting, VRADA can attain an $\mathcal{O}(\frac{1}{n})$ -accurate solution in $\mathcal{O}(n \log \log n)$ number of stochastic gradient evaluations, where n is the number of samples; meanwhile, VRADA matches the lower bound of this setting up to a $\log \log n$ factor. In the strongly convex and smooth setting, VRADA matches the lower bound in the regime $n \leq \Theta(\kappa)$, while it improves the rate in the regime $n \geq \kappa$ to $\mathcal{O}(n + \frac{n \log(1/\epsilon)}{\log(n/\kappa)})$, where κ is the condition number. Besides improving the best known complexity results, VRADA has more unified and simplified algorithmic implementation and convergence analysis for both the general convex and strongly convex settings. Through experiments on real datasets, we show the good performance of VRADA over existing methods for large-scale machine learning problems.

[Tree! I am no Tree! I am a low dimensional Hyperbolic Embedding](#)

- Rishi Sonthalia, Anna Gilbert
- abstract@[open-review](#): Given data, finding a faithful low-dimensional hyperbolic embedding of the data is a key method by which we can extract hierarchical information or learn representative geometric features of the data. In this paper, we explore a new method for learning hyperbolic representations by taking a metric-first approach. Rather than determining the low-dimensional hyperbolic embedding directly, we learn a tree structure on the data. This tree structure can then be used directly to extract hierarchical information, embedded into a hyperbolic manifold using Sarkar's construction \textit{sarkar}, or used as a tree approximation of the original metric. To this end, we present a novel fast algorithm \textit{TreeRep} such that, given a δ -hyperbolic metric (for any $\delta \geq 0$), the algorithm learns a tree structure that approximates the original metric. In the case when $\delta = 0$, we show analytically that \textit{TreeRep} exactly recovers the original tree structure. We show empirically that \textit{TreeRep} is not only many orders of magnitude faster than previously known algorithms, but also produces metrics with lower average distortion and higher mean average precision than most previous algorithms for learning hyperbolic embeddings, extracting hierarchical information, and approximating metrics via tree metrics.

[Deep Structural Causal Models for Tractable Counterfactual Inference](#)

- Nick Pawlowski, Daniel Coelho de Castro, Ben Glocker
- abstract@[open-review](#): We formulate a general framework for building structural causal models (SCMs) with deep learning components. The proposed approach employs normalising flows and variational inference to enable tractable inference of exogenous noise variables - a crucial step for counterfactual inference that is missing from existing deep causal learning methods. Our framework is validated on a synthetic dataset built on MNIST as well as on a real-world medical dataset of brain MRI scans. Our experimental results indicate that we can successfully train deep SCMs that are capable of all three levels of Pearl's ladder of causation: association, intervention, and counterfactuals, giving rise to a powerful new approach for answering causal questions in imaging applications and beyond.

[Convolutional Generation of Textured 3D Meshes](#)

- Dario Pavllo, Graham Spinks, Thomas Hofmann, Marie-Francine Moens, Aurelien Lucchi

- abstract@[open-review](#): While recent generative models for 2D images achieve impressive visual results, they clearly lack the ability to perform 3D reasoning. This heavily restricts the degree of control over generated objects as well as the possible applications of such models. In this work, we bridge this gap by leveraging recent advances in differentiable rendering. We design a framework that can generate triangle meshes and associated high-resolution texture maps, using only 2D supervision from single-view natural images. A key contribution of our work is the encoding of the mesh and texture as 2D representations, which are semantically aligned and can be easily modeled by a 2D convolutional GAN. We demonstrate the efficacy of our method on Pascal3D+ Cars and CUB, both in an unconditional setting and in settings where the model is conditioned on class labels, attributes, and text. Finally, we propose an evaluation methodology that assesses the mesh and texture quality separately.

[A Statistical Framework for Low-bitwidth Training of Deep Neural Networks](#)

- Jianfei Chen, Yu Gai, Zhewei Yao, Michael W. Mahoney, Joseph E. Gonzalez
- abstract@[open-review](#): Fully quantized training (FQT), which uses low-bitwidth hardware by quantizing the activations, weights, and gradients of a neural network model, is a promising approach to accelerate the training of deep neural networks. One major challenge with FQT is the lack of theoretical understanding, in particular of how gradient quantization impacts convergence properties. In this paper, we address this problem by presenting a statistical framework for analyzing FQT algorithms. We view the quantized gradient of FQT as a stochastic estimator of its full precision counterpart, a procedure known as quantization-aware training (QAT). We show that the FQT gradient is an unbiased estimator of the QAT gradient, and we discuss the impact of gradient quantization on its variance. Inspired by these theoretical results, we develop two novel gradient quantizers, and we show that these have smaller variance than the existing per-tensor quantizer. For training ResNet-50 on ImageNet, our 5-bit block Householder quantizer achieves only 0.5% validation accuracy loss relative to QAT, comparable to the existing INT8 baseline.

[Better Set Representations For Relational Reasoning](#)

- Qian Huang, Horace He, Abhay Singh, Yan Zhang, Ser Nam Lim, Austin R. Benson
- abstract@[open-review](#): Incorporating relational reasoning into neural networks has greatly expanded their capabilities and scope. One defining trait of relational reasoning is that it operates on a set of entities, as opposed to standard vector representations. Existing end-to-end approaches for relational reasoning typically extract entities from inputs by directly interpreting the latent feature representations as a set. We show that these approaches do not respect set permutational invariance and thus have fundamental representational limitations. To resolve this limitation, we propose a simple and general network module called Set Refiner Network (SRN). We first use synthetic image experiments to demonstrate how our approach effectively decomposes objects without explicit supervision. Then, we insert our module into existing relational reasoning models and show that respecting set invariance leads to substantial gains in prediction performance and robustness on several relational reasoning tasks. Code can be found at github.com/CUAI/BetterSetRepresentations.

[AutoSync: Learning to Synchronize for Data-Parallel Distributed Deep Learning](#)

- Hao Zhang, Yuan Li, Zijie Deng, Xiaodan Liang, Lawrence Carin, Eric Xing
- abstract@[open-review](#): Synchronization is a key step in data-parallel distributed machine learning (ML). Different synchronization systems and strategies perform differently, and to achieve optimal parallel training throughput requires synchronization strategies that adapt to model structures and cluster configurations. Existing synchronization systems often only consider a single or a few synchronization aspects, and the burden of deciding the right synchronization strategy is then placed on the ML practitioners, who may lack the required expertise. In this paper, we develop a model- and resource-dependent representation for synchronization, which unifies multiple synchronization aspects ranging from architecture, message partitioning, placement scheme, to communication topology. Based on this representation, we build an end-to-end pipeline, AutoSync, to automatically optimize synchronization strategies given model structures and resource specifications, lowering the bar for data-parallel distributed ML. By learning from low-shot data collected in only 200 trial runs, AutoSync can discover synchronization strategies up to 1.6x better than manually optimized ones. We develop transfer-learning mechanisms to further reduce the auto-optimization cost -- the simulators can transfer among similar model architectures, among similar cluster configurations, or both. We also present a dataset that contains over 10000 synchronization strategies and run-time pairs on a diverse set of models and cluster specifications.

[A Combinatorial Perspective on Transfer Learning](#)

- Jianan Wang, Eren Sezener, David Budden, Marcus Hutter, Joel Veness
- abstract@[open-review](#): Human intelligence is characterized not only by the capacity to learn complex skills, but the ability to rapidly adapt and acquire new skills within an ever-changing environment. In this work we study how the learning of modular solutions can allow for effective generalization to both unseen and potentially differently distributed data. Our main postulate is that the combination of task segmentation, modular learning and memory-based ensembling can give rise to generalization on an exponentially growing number of unseen tasks. We provide a concrete instantiation of this idea using a combination of: (1) the Forget-Me-Not Process, for task segmentation and memory based ensembling; and (2) Gated Linear Networks, which in contrast to contemporary deep learning techniques use a modular and local learning mechanism. We demonstrate that this system exhibits a number of desirable continual learning properties: robustness to catastrophic forgetting, no negative transfer and increasing levels of positive transfer as more tasks are seen. We show competitive performance against both offline and online methods on standard continual learning benchmarks.

[Hardness of Learning Neural Networks with Natural Weights](#)

- Amit Daniely, Gal Vardi
- abstract@[open-review](#): Neural networks are nowadays highly successful despite strong hardness results. The existing hardness results focus on the network architecture, and assume that the network's weights are arbitrary. A natural approach to settle the discrepancy is to assume that the network's weights are "well-behaved" and possess some generic properties that may allow efficient learning. This approach is supported by the intuition that the weights in real-world networks are not arbitrary, but exhibit some "random-like" properties with respect to some "natural" distributions. We prove negative results in this regard, and show that for depth-\$2\$ networks, and many natural" weights distributions such as the normal and the uniform distribution, most networks are hard to learn. Namely, there is no efficient learning algorithm that is provably successful for most weights, and every input distribution. It implies that there is no generic property that holds with high probability in such random networks and allows efficient learning.

[Higher-Order Spectral Clustering of Directed Graphs](#)

- Steinar Laenen, He Sun
- abstract@[open-review](#): Clustering is an important topic in algorithms, and has a number of applications in machine learning, computer vision, statistics, and several other research disciplines. Traditional objectives of graph clustering are to find clusters with low conductance. Not only are these objectives just applicable for undirected graphs, they are also incapable to take the relationships between clusters into account, which could be crucial for many applications. To overcome these downsides, we study directed graphs (digraphs) whose clusters exhibit further "structural" information amongst each other. Based on the Hermitian matrix representation of digraphs, we present a nearly-linear time algorithm for digraph clustering, and further show that our proposed algorithm can be implemented in sublinear time under reasonable assumptions. The significance of our theoretical work is demonstrated by extensive experimental results on the UN Comtrade Dataset: the output clustering of our algorithm exhibits not only how the clusters (sets of countries)

relate to each other with respect to their import and export records, but also how these clusters evolve over time, in accordance with known facts in international trade.

[Primal-Dual Mesh Convolutional Neural Networks](#)

- Francesco Milano, Antonio Loquercio, Antoni Rosinol, Davide Scaramuzza, Luca Carlone
- abstract@[open-review](#): Recent works in geometric deep learning have introduced neural networks that allow performing inference tasks on three-dimensional geometric data by defining convolution --and sometimes pooling-- operations on triangle meshes. These methods, however, either consider the input mesh as a graph, and do not exploit specific geometric properties of meshes for feature aggregation and downsampling, or are specialized for meshes, but rely on a rigid definition of convolution that does not properly capture the local topology of the mesh. We propose a method that combines the advantages of both types of approaches, while addressing their limitations: we extend a primal-dual framework drawn from the graph-neural-network literature to triangle meshes, and define convolutions on two types of graphs constructed from an input mesh. Our method takes features for both edges and faces of a 3D mesh as input, and dynamically aggregates them using an attention mechanism. At the same time, we introduce a pooling operation with a precise geometric interpretation, that allows handling variations in the mesh connectivity by clustering mesh faces in a task-driven fashion. We provide theoretical insights of our approach using tools from the mesh-simplification literature. In addition, we validate experimentally our method in the tasks of shape classification and shape segmentation, where we obtain comparable or superior performance to the state of the art.

[The Advantage of Conditional Meta-Learning for Biased Regularization and Fine Tuning](#)

- Giulia Denevi, Massimiliano Pontil, Carlo Ciliberto
- abstract@[open-review](#): Biased regularization and fine tuning are two recent meta-learning approaches. They have been shown to be effective to tackle distributions of tasks, in which the tasks' target vectors are all close to a common meta-parameter vector. However, these methods may perform poorly on heterogeneous environments of tasks, where the complexity of the tasks' distribution cannot be captured by a single meta-parameter vector. We address this limitation by conditional meta-learning, inferring a conditioning function mapping task's side information into a meta-parameter vector that is appropriate for that task at hand. We characterize properties of the environment under which the conditional approach brings a substantial advantage over standard meta-learning and we highlight examples of environments, such as those with multiple clusters, satisfying these properties. We then propose a convex meta-algorithm providing a comparable advantage also in practice. Numerical experiments confirm our theoretical findings.

[Watch out! Motion is Blurring the Vision of Your Deep Neural Networks](#)

- Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Jian Wang, Bing Yu, Wei Feng, Yang Liu
- abstract@[open-review](#): The state-of-the-art deep neural networks (DNNs) are vulnerable against adversarial examples with additive random-like noise perturbations. While such examples are hardly found in the physical world, the image blurring effect caused by object motion, on the other hand, commonly occurs in practice, making the study of which greatly important especially for the widely adopted real-time image processing tasks (e.g., object detection, tracking). In this paper, we initiate the first step to comprehensively investigate the potential hazards of blur effect for DNN, caused by object motion. We propose a novel adversarial attack method that can generate visually natural motion-blurred adversarial examples, named motion-based adversarial blur attack (ABBA). To this end, we first formulate the kernel-prediction-based attack where an input image is convolved with kernels in a pixel-wise way, and the misclassification capability is achieved by tuning the kernel weights. To generate visually more natural and plausible examples, we further propose the saliency-regularized adversarial kernel prediction, where the salient region serves as a moving object, and the predicted kernel is regularized to achieve naturally visual effects. Besides, the attack is further enhanced by adaptively tuning the translations of object and background. A comprehensive evaluation on the NeurIPS'17 adversarial competition dataset demonstrates the effectiveness of ABBA by considering various kernel sizes, translations, and regions. The in-depth study further confirms that our method shows a more effective penetrating capability to the state-of-the-art GAN-based deblurring mechanisms compared with other blurring methods. We release the code to \url{https://github.com/tsingqguo/ABBA}.

[Sinkhorn Barycenter via Functional Gradient Descent](#)

- Zebang Shen, Zhenfu Wang, Alejandro Ribeiro, Hamed Hassani
- abstract@[open-review](#): In this paper, we consider the problem of computing the barycenter of a set of probability distributions under the Sinkhorn divergence. This problem has recently found applications across various domains, including graphics, learning, and vision, as it provides a meaningful mechanism to aggregate knowledge. Unlike previous approaches which directly operate in the space of probability measures, we recast the Sinkhorn barycenter problem as an instance of unconstrained functional optimization and develop a novel functional gradient descent method named \texttt{Sinkhorn Descent} (\texttt{SD}). We prove that \texttt{SD} converges to a stationary point at a sublinear rate, and under reasonable assumptions, we further show that it asymptotically finds a global minimizer of the Sinkhorn barycenter problem. Moreover, by providing a mean-field analysis, we show that \texttt{SD} preserves the \{weak convergence\} of empirical measures. Importantly, the computational complexity of \texttt{SD} scales linearly in the dimension \$d\$ and we demonstrate its scalability by solving a \$100\$-dimensional Sinkhorn barycenter problem.

[Coresets for Near-Convex Functions](#)

- Murad Tukan, Alaa Maalouf, Dan Feldman
- abstract@[open-review](#): Coreset is usually a small weighted subset of \$n\$ input points in \$\mathbb{R}^d\$, that provably approximates their loss function for a given set of queries (models, classifiers, etc.). Coresets become increasingly common in machine learning since existing heuristics or inefficient algorithms may be improved by running them possibly many times on the small coreset that can be maintained for streaming distributed data. Coresets can be obtained by sensitivity (importance) sampling, where its size is proportional to the total sum of sensitivities. Unfortunately, computing the sensitivity of each point is problem dependent and may be harder to compute than the original optimization problem at hand. We suggest a generic framework for computing sensitivities (and thus coresets) for wide family of loss functions which we call near-convex functions. This is by suggesting the \$f\$-SVD factorization that generalizes the SVD factorization of matrices to functions. Example applications include coresets that are either new or significantly improves previous results, such as SVM, Logistic regression, M-estimators, and \$\ell_1\$-regression. Experimental results and open source are also provided.

[Bayesian Deep Ensembles via the Neural Tangent Kernel](#)

- Bobby He, Balaji Lakshminarayanan, Yee Whye Teh
- abstract@[open-review](#): We explore the link between deep ensembles and Gaussian processes (GPs) through the lens of the Neural Tangent Kernel (NTK): a recent development in understanding the training dynamics of wide neural networks (NNs). Previous work has shown that even in the infinite width limit, when NNs become GPs, there is no GP posterior interpretation to a deep ensemble trained with squared error loss. We introduce a simple modification to standard deep ensembles training, through addition of a computationally-tractable, randomised and untrainable function to each ensemble member, that enables a posterior interpretation in the infinite width limit. When ensembled together, our trained NNs give an approximation to a posterior predictive distribution, and we prove that our Bayesian deep ensembles make more conservative predictions than standard deep ensembles in the infinite width limit. Finally, using finite width NNs we demonstrate that our Bayesian deep ensembles faithfully emulate the analytic posterior predictive when available, and outperform standard deep ensembles in various out-of-distribution settings, for both regression and classification tasks.

Improved Schemes for Episodic Memory-based Lifelong Learning

- Yunhui Guo, Mingrui Liu, Tianbao Yang, Tajana Rosing
- abstract@[open-review](#): Current deep neural networks can achieve remarkable performance on a single task. However, when the deep neural network is continually trained on a sequence of tasks, it seems to gradually forget the previous learned knowledge. This phenomenon is referred to as catastrophic forgetting and motivates the field called lifelong learning. Recently, episodic memory based approaches such as GEM and A-GEM have shown remarkable performance. In this paper, we provide the first unified view of episodic memory based approaches from an optimization's perspective. This view leads to two improved schemes for episodic memory based lifelong learning, called MEGA- \setminus rom{1} and MEGA- \setminus rom{2}. MEGA- \setminus rom{1} and MEGA- \setminus rom{2} modulate the balance between old tasks and the new task by integrating the current gradient with the gradient computed on the episodic memory. Notably, we show that GEM and A-GEM are degenerate cases of MEGA- \setminus rom{1} and MEGA- \setminus rom{2} which consistently put the same emphasis on the current task, regardless of how the loss changes over time. Our proposed schemes address this issue by using novel loss-balancing updating rules, which drastically improve the performance over GEM and A-GEM. Extensive experimental results show that the proposed schemes significantly advance the state-of-the-art on four commonly used lifelong learning benchmarks, reducing the error by up to 18%.

Adaptive Sampling for Stochastic Risk-Averse Learning

- Sebastian Curi, Kfir Y. Levy, Stefanie Jegelka, Andreas Krause
- abstract@[open-review](#): In high-stakes machine learning applications, it is crucial to not only perform well {em on average}, but also when restricted to {em difficult} examples. To address this, we consider the problem of training models in a risk-averse manner. We propose an adaptive sampling algorithm for stochastically optimizing the {em Conditional Value-at-Risk (CVaR)} of a loss distribution, which measures its performance on the α fraction of most difficult examples. We use a distributionally robust formulation of the CVaR to phrase the problem as a zero-sum game between two players, and solve it efficiently using regret minimization. Our approach relies on sampling from structured Determinantal Point Processes (DPPs), which enables scaling it to large data sets. Finally, we empirically demonstrate its effectiveness on large-scale convex and non-convex learning tasks.

Deep Wiener Deconvolution: Wiener Meets Deep Learning for Image Deblurring

- Jiangxin Dong, Stefan Roth, Bernt Schiele
- abstract@[open-review](#): We present a simple and effective approach for non-blind image deblurring, combining classical techniques and deep learning. In contrast to existing methods that deblur the image directly in the standard image space, we propose to perform an explicit deconvolution process in a feature space by integrating a classical Wiener deconvolution framework with learned deep features. A multi-scale feature refinement module then predicts the deblurred image from the deconvolved deep features, progressively recovering detail and small-scale structures. The proposed model is trained in an end-to-end manner and evaluated on scenarios with both simulated and real-world image blur. Our extensive experimental results show that the proposed deep Wiener deconvolution network facilitates deblurred results with visibly fewer artifacts. Moreover, our approach quantitatively outperforms state-of-the-art non-blind image deblurring methods by a wide margin.

Discovering Reinforcement Learning Algorithms

- Junhyuk Oh, Matteo Hessel, Wojciech M. Czarnecki, Zhongwen Xu, Hado P. van Hasselt, Satinder Singh, David Silver
- abstract@[open-review](#): Reinforcement learning (RL) algorithms update an agent's parameters according to one of several possible rules, discovered manually through years of research. Automating the discovery of update rules from data could lead to more efficient algorithms, or algorithms that are better adapted to specific environments. Although there have been prior attempts at addressing this significant scientific challenge, it remains an open question whether it is feasible to discover alternatives to fundamental concepts of RL such as value functions and temporal-difference learning. This paper introduces a new meta-learning approach that discovers an entire update rule which includes both what to predict' (e.g. value functions) and how to learn from it' (e.g. bootstrapping) by interacting with a set of environments. The output of this method is an RL algorithm that we call Learned Policy Gradient (LPG). Empirical results show that our method discovers its own alternative to the concept of value functions. Furthermore it discovers a bootstrapping mechanism to maintain and use its predictions. Surprisingly, when trained solely on toy environments, LPG generalises effectively to complex Atari games and achieves non-trivial performance. This shows the potential to discover general RL algorithms from data.

Taming Discrete Integration via the Boon of Dimensionality

- Jeffrey Dudek, Dror Fried, Kuldeep S Meel
- abstract@[open-review](#): Discrete integration is a fundamental problem in computer science that concerns the computation of discrete sums over exponentially large sets. Despite intense interest from researchers for over three decades, the design of scalable techniques for computing estimates with rigorous guarantees for discrete integration remains the holy grail. The key contribution of this work addresses this scalability challenge via an efficient reduction of discrete integration to model counting. The proposed reduction is achieved via a significant increase in the dimensionality that, contrary to conventional wisdom, leads to solving an instance of the relatively simpler problem of model counting. Building on the promising approach proposed by Chakraborty et al, our work overcomes the key weakness of their approach: a restriction to dyadic weights. We augment our proposed reduction, called DeWeight, with a state of the art efficient approximate model counter and perform detailed empirical analysis over benchmarks arising from neural network verification domains, an emerging application area of critical importance. DeWeight, to the best of our knowledge, is the first technique to compute estimates with provable guarantees for this class of benchmarks.

Blind Video Temporal Consistency via Deep Video Prior

- Chenyang Lei, Yazhou Xing, Qifeng Chen
- abstract@[open-review](#): Applying image processing algorithms independently to each video frame often leads to temporal inconsistency in the resulting video. To address this issue, we present a novel and general approach for blind video temporal consistency. Our method is only trained on a pair of original and processed videos directly instead of a large dataset. Unlike most previous methods that enforce temporal consistency with optical flow, we show that temporal consistency can be achieved by training a convolutional network on a video with the Deep Video Prior. Moreover, a carefully designed iteratively reweighted training strategy is proposed to address the challenging multimodal inconsistency problem. We demonstrate the effectiveness of our approach on 7 computer vision tasks on videos. Extensive quantitative and perceptual experiments show that our approach obtains superior performance than state-of-the-art methods on blind video temporal consistency.

Simplify and Robustify Negative Sampling for Implicit Collaborative Filtering

- Jingtao Ding, Yuhan Quan, Quanming Yao, Yong Li, Depeng Jin
- abstract@[open-review](#): Negative sampling approaches are prevalent in implicit collaborative filtering for obtaining negative labels from massive unlabeled data. As two major concerns in negative sampling, efficiency and effectiveness are still not fully achieved by recent works that use complicate structures and overlook risk of false negative instances. In this paper, we first provide a novel understanding of negative instances by empirically observing that only a few instances are potentially important for model learning, and false negatives tend to have stable predictions over many training iterations. Above findings motivate us to simplify the model by sampling from designed memory that only stores a few important candidates

and, more importantly, tackle the untouched false negative problem by favouring high-variance samples stored in memory, which achieves efficient sampling of true negatives with high-quality. Empirical results on two synthetic datasets and three real-world datasets demonstrate both robustness and superiorities of our negative sampling method. The implementation is available at <https://github.com/dingjingtao/SRNS>.

[Model Selection for Production System via Automated Online Experiments](#)

- Zhenwen Dai, Praveen Chandar, Ghazal Fazelnia, Benjamin Carterette, Mounia Lalmas
- abstract@[open-review](#): A challenge that machine learning practitioners in the industry face is the task of selecting the best model to deploy in production. As a model is often an intermediate component of a production system, online controlled experiments such as A/B tests yield the most reliable estimation of the effectiveness of the whole system, but can only compare two or a few models due to budget constraints. We propose an automated online experimentation mechanism that can efficiently perform model selection from a large pool of models with a small number of online experiments. We derive the probability distribution of the metric of interest that contains the model uncertainty from our Bayesian surrogate model trained using historical logs. Our method efficiently identifies the best model by sequentially selecting and deploying a list of models from the candidate set that balance exploration-exploitation. Using simulations based on real data, we demonstrate the effectiveness of our method on two different tasks.

[On the Almost Sure Convergence of Stochastic Gradient Descent in Non-Convex Problems](#)

- Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis, Volkan Cevher
- abstract@[open-review](#): In this paper, we analyze the trajectories of stochastic gradient descent (SGD) with the aim of understanding their convergence properties in non-convex problems. We first show that the sequence of iterates generated by SGD remains bounded and converges with probability 1 under a very broad range of step-size schedules. Subsequently, we prove that the algorithm's rate of convergence to local minimizers with a positive-definite Hessian is $O(1/n^p)$ if the method is run with a $(1/n^p)$ step-size. This provides an important guideline for tuning the algorithm's step-size as it suggests that a cool-down phase with a vanishing step-size could lead to significant performance gains; we demonstrate this heuristic using ResNet architectures on CIFAR. Finally, going beyond existing positive probability guarantees, we show that SGD avoids strict saddle points/manifolds with probability 1 for the entire spectrum of step-size policies considered.

[Automatic Perturbation Analysis for Scalable Certified Robustness and Beyond](#)

- Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, Cho-Jui Hsieh
- abstract@[open-review](#): Linear relaxation based perturbation analysis (LiRPA) for neural networks, which computes provable linear bounds of output neurons given a certain amount of input perturbation, has become a core component in robustness verification and certified defense. The majority of LiRPA-based methods focus on simple feed-forward networks and need particular manual derivations and implementations when extended to other architectures. In this paper, we develop an automatic framework to enable perturbation analysis on any neural network structures, by generalizing existing LiRPA algorithms such as CROWN to operate on general computational graphs. The flexibility, differentiability and ease of use of our framework allow us to obtain state-of-the-art results on LiRPA based certified defense on fairly complicated networks like DenseNet, ResNeXt and Transformer that are not supported by prior works. Our framework also enables loss fusion, a technique that significantly reduces the computational complexity of LiRPA for certified defense. For the first time, we demonstrate LiRPA based certified defense on Tiny ImageNet and Downscaled ImageNet where previous approaches cannot scale to due to the relatively large number of classes. Our work also yields an open-source library for the community to apply LiRPA to areas beyond certified defense without much LiRPA expertise, e.g., we create a neural network with a provably flat optimization landscape by applying LiRPA to network parameters. Our open source library is available at https://github.com/KaidiXu/auto_LiRPA.

[Adaptation Properties Allow Identification of Optimized Neural Codes](#)

- Luke Rast, Jan Drugowitsch
- abstract@[open-review](#): The adaptation of neural codes to the statistics of their environment is well captured by efficient coding approaches. Here we solve an inverse problem: characterizing the objective and constraint functions that efficient codes appear to be optimal for, on the basis of how they adapt to different stimulus distributions. We formulate a general efficient coding problem, with flexible objective and constraint functions and minimal parametric assumptions. Solving special cases of this model, we provide solutions to broad classes of Fisher information-based efficient coding problems, generalizing a wide range of previous results. We show that different objective function types impose qualitatively different adaptation behaviors, while constraints enforce characteristic deviations from classic efficient coding signatures. Despite interaction between these effects, clear signatures emerge for both unconstrained optimization problems and information-maximizing objective functions. Asking for a fixed-point of the neural code adaptation, we find an objective-independent characterization of constraints on the neural code. We use this result to propose an experimental paradigm that can characterize both the objective and constraint functions that an observed code appears to be optimized for.

[Global Convergence and Variance Reduction for a Class of Nonconvex-Nonconcave Minimax Problems](#)

- Junchi Yang, Negar Kiyavash, Niao He
- abstract@[open-review](#): Nonconvex minimax problems appear frequently in emerging machine learning applications, such as generative adversarial networks and adversarial learning. Simple algorithms such as the gradient descent ascent (GDA) are the common practice for solving these nonconvex games and receive lots of empirical success. Yet, it is known that these vanilla GDA algorithms with constant stepsize can potentially diverge even in the convex setting. In this work, we show that for a subclass of nonconvex-nonconcave objectives satisfying a so-called two-sided Polyak-{L}ojasiewicz inequality, the alternating gradient descent ascent (AGDA) algorithm converges globally at a linear rate and the stochastic AGDA achieves a sublinear rate. We further develop a variance reduced algorithm that attains a provably faster rate than AGDA when the problem has the finite-sum structure.

[Model-Based Multi-Agent RL in Zero-Sum Markov Games with Near-Optimal Sample Complexity](#)

- Kaiqing Zhang, Sham Kakade, Tamer Basar, Lin Yang
- abstract@[open-review](#): Model-based reinforcement learning (RL), which finds an optimal policy using an empirical model, has long been recognized as one of the cornerstones of RL. It is especially suitable for multi-agent RL (MARL), as it naturally decouples the learning and the planning phases, and avoids the non-stationarity problem when all agents are improving their policies simultaneously using samples. Though intuitive and widely-used, the sample complexity of model-based MARL algorithms has been investigated relatively much less often. In this paper, we aim to address the fundamental open question about the sample complexity of model-based MARL. We study arguably the most basic MARL setting: two-player discounted zero-sum Markov games, given only access to a generative model of state transition. We show that model-based MARL achieves a sample complexity of $\tilde{O}(\|cS\|\|cA\|\|cB\|(1-\gamma)^{-3}\epsilon^{-2})$ for finding the Nash equilibrium (NE) up to some ϵ error, and the ϵ -NE policies, where γ is the discount factor, and cS, cA, cB denote the state space, and the action spaces for the two agents. We also show that this method is near-minimax optimal with a tight dependence on $1-\gamma$ and $\|cS\|$ by providing a lower bound of $\Omega(\|cS\|(\|cA\| + \|cB\|)(1-\gamma)^{-3}\epsilon^{-2})$. Our results justify the efficiency of this simple model-based approach in the multi-agent RL setting.

[Conservative Q-Learning for Offline Reinforcement Learning](#)

- Aviral Kumar, Aurick Zhou, George Tucker, Sergey Levine

- abstract@[open-review](#): Effectively leveraging large, previously collected datasets in reinforcement learning (RL) is a key challenge for large-scale real-world applications. Offline RL algorithms promise to learn effective policies from previously-collected, static datasets without further interaction. However, in practice, offline RL presents a major challenge, and standard off-policy RL methods can fail due to overestimation of values induced by the distributional shift between the dataset and the learned policy, especially when training on complex and multi-modal data distributions. In this paper, we propose conservative Q-learning (CQL), which aims to address these limitations by learning a conservative Q-function such that the expected value of a policy under this Q-function lower-bounds its true value. We theoretically show that CQL produces a lower bound on the value of the current policy and that it can be incorporated into a policy learning procedure with theoretical improvement guarantees. In practice, CQL augments the standard Bellman error objective with a simple Q-value regularizer which is straightforward to implement on top of existing deep Q-learning and actor-critic implementations. On both discrete and continuous control domains, we show that CQL substantially outperforms existing offline RL methods, often learning policies that attain 2-5 times higher final return, especially when learning from complex and multi-modal data distributions.

Online Influence Maximization under Linear Threshold Model

- Shuai Li, Fang Kong, Kejie Tang, Qizhi Li, Wei Chen
- abstract@[open-review](#): Online influence maximization (OIM) is a popular problem in social networks to learn influence propagation model parameters and maximize the influence spread at the same time. Most previous studies focus on the independent cascade (IC) model under the edge-level feedback. In this paper, we address OIM in the linear threshold (LT) model. Because node activations in the LT model are due to the aggregated effect of all active neighbors, it is more natural to model OIM with the node-level feedback. And this brings new challenge in online learning since we only observe aggregated effect from groups of nodes and the groups are also random. Based on the linear structure in node activations, we incorporate ideas from linear bandits and design an algorithm \$ltlinucb\$ that is consistent with the observed feedback. By proving group observation modulated (GOM) bounded smoothness property, a novel result of the influence difference in terms of the random observations, we provide a regret of order \$\tilde{O}(\mathrm{poly}(m)\sqrt{T})\$, where \$m\$ is the number of edges and \$T\$ is the number of rounds. This is the first theoretical result in such order for OIM under the LT model. In the end, we also provide an algorithm \$oimetc\$ with regret bound \$O(\mathrm{poly}(m)T^{2/3})\$, which is model-independent, simple and has less requirement on online feedback and offline computation.

Ensembling geophysical models with Bayesian Neural Networks

- Ushnish Sengupta, Matt Amos, Scott Hosking, Carl Edward Rasmussen, Matthew Juniper, Paul Young
- abstract@[open-review](#): Ensembles of geophysical models improve projection accuracy and express uncertainties. We develop a novel data-driven ensembling strategy for combining geophysical models using Bayesian Neural Networks, which infers spatiotemporally varying model weights and bias while accounting for heteroscedastic uncertainties in the observations. This produces more accurate and uncertainty-aware projections without sacrificing interpretability. Applied to the prediction of total column ozone from an ensemble of 15 chemistry-climate models, we find that the Bayesian neural network ensemble (BayNNE) outperforms existing ensembling methods, achieving a 49.4% reduction in RMSE for temporal extrapolation, and a 67.4% reduction in RMSE for polar data voids, compared to a weighted mean. Uncertainty is also well-characterized, with 90.6% of the data points in our extrapolation validation dataset lying within 2 standard deviations and 98.5% within 3 standard deviations.

Delving into the Cyclic Mechanism in Semi-supervised Video Object Segmentation

- Yuxi Li, Ning Xu, Jinlong Peng, John See, Weiyao Lin
- abstract@[open-review](#): In this paper, we take attempt to incorporate the cyclic mechanism with the vision task of semi-supervised video object segmentation. By resorting to the accurate reference mask of the first frame, we try to mitigate the error propagation problem in most of current video object segmentation pipelines. Firstly, we propose a cyclic scheme for offline training of segmentation networks. Then, we extend the offline pipeline to an online method by introducing a simple gradient correction module while keeping high efficiency as other offline methods. Finally we develop cycle effective receptive field (cycle-ERF) from gradient correction to provide a new perspective for analyzing object-specific regions of interests. We conduct comprehensive experiments on benchmarks of DAVIS17 and Youtube-VOS, demonstrating that our introduced cyclic mechanism is helpful to boost the segmentation quality.

Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability

- Christopher Frye, Colin Rowat, Ilya Feige
- abstract@[open-review](#): Explaining AI systems is fundamental both to the development of high performing models and to the trust placed in them by their users. The Shapley framework for explainability has strength in its general applicability combined with its precise, rigorous foundation: it provides a common, model-agnostic language for AI explainability and uniquely satisfies a set of intuitive mathematical axioms. However, Shapley values are too restrictive in one significant regard: they ignore all causal structure in the data. We introduce a less restrictive framework, Asymmetric Shapley values (ASVs), which are rigorously founded on a set of axioms, applicable to any AI system, and can flexibly incorporate any causal structure known to be respected by the data. We demonstrate that ASVs can (i) improve model explanations by incorporating causal information, (ii) provide an unambiguous test for unfair discrimination in model predictions, (iii) enable sequentially incremental explanations in time-series models, and (iv) support feature-selection studies without the need for model retraining.

Understanding Deep Architecture with Reasoning Layer

- Xinshi Chen, Yufei Zhang, Christoph Reisinger, Le Song
- abstract@[open-review](#): Recently, there is a surge of interest in combining deep learning models with reasoning in order to handle more sophisticated learning tasks. In many cases, a reasoning task can be solved by an iterative algorithm. This algorithm is often unrolled, truncated, and used as a specialized layer in the deep architecture, which can be trained end-to-end with other neural components. Although such hybrid deep architectures have led to many empirical successes, theoretical understandings of such architectures, especially the interplay between algorithm layers and other neural layers, remains largely unexplored. In this paper, we take an initial step toward an understanding of such hybrid deep architectures by showing that properties of the algorithm layers, such as convergence, stability and sensitivity, are intimately related to the approximation and generalization abilities of the end-to-end model. Furthermore, our analysis matches nicely with experimental observations under various conditions, suggesting that our theory can provide useful guidelines for designing deep architectures with reasoning layers.

Planning in Markov Decision Processes with Gap-Dependent Sample Complexity

- Anders Jonsson, Emilie Kaufmann, Pierre Menard, Omar Darwiche Domingues, Edouard Leurent, Michal Valko
- abstract@[open-review](#): We propose MDP-GapE, a new trajectory-based Monte-Carlo Tree Search algorithm for planning in a Markov Decision Process in which transitions have a finite support. We prove an upper bound on the number of sampled trajectories needed for MDP-GapE to identify a near-optimal action with high probability. This problem-dependent result is expressed in terms of the sub-optimality gaps of the state-action pairs that are visited during exploration. Our experiments reveal that MDP-GapE is also effective in practice, in contrast with other algorithms with sample complexity guarantees in the fixed-confidence setting, that are mostly theoretical.

Provably Good Batch Off-Policy Reinforcement Learning Without Great Exploration

- Yao Liu, Adith Swaminathan, Alekh Agarwal, Emma Brunskill
- abstract@[open-review](#): Batch reinforcement learning (RL) is important to apply RL algorithms to many high stakes tasks. Doing batch RL in a way that yields a reliable new policy in large domains is challenging: a new decision policy may visit states and actions outside the support of the batch data, and function approximation and optimization with limited samples can further increase the potential of learning policies with overly optimistic estimates of their future performance. Some recent approaches to address these concerns have shown promise, but can still be overly optimistic in their expected outcomes. Theoretical work that provides strong guarantees on the performance of the output policy relies on a strong concentrability assumption, which makes it unsuitable for cases where the ratio between state-action distributions of behavior policy and some candidate policies is large. This is because, in the traditional analysis, the error bound scales up with this ratio. We show that using \emph{pessimistic value estimates} in the low-data regions in Bellman optimality and evaluation back-up can yield more adaptive and stronger guarantees when the concentrability assumption does not hold. In certain settings, they can find the approximately best policy within the state-action space explored by the batch data, without requiring a priori assumptions of concentrability. We highlight the necessity of our pessimistic update and the limitations of previous algorithms and analyses by illustrative MDP examples and demonstrate an empirical comparison of our algorithm and other state-of-the-art batch RL baselines in standard benchmarks.

[Detection as Regression: Certified Object Detection with Median Smoothing](#)

- Ping-yeh Chiang, Michael Curry, Ahmed Abdelkader, Aounon Kumar, John Dickerson, Tom Goldstein
- abstract@[open-review](#): Despite the vulnerability of object detectors to adversarial attacks, very few defenses are known to date. While adversarial training can improve the empirical robustness of image classifiers, a direct extension to object detection is very expensive. This work is motivated by recent progress on certified classification by randomized smoothing. We start by presenting a reduction from object detection to a regression problem. Then, to enable certified regression, where standard mean smoothing fails, we propose median smoothing, which is of independent interest. We obtain the first model-agnostic, training-free, and certified defense for object detection against ℓ_2 -bounded attacks.

[Contextual Reserve Price Optimization in Auctions via Mixed Integer Programming](#)

- Joey Huchette, Haihao Lu, Hossein Esfandiari, Vahab Mirrokni
- abstract@[open-review](#): We study the problem of learning a linear model to set the reserve price in an auction, given contextual information, in order to maximize expected revenue from the seller side. First, we show that it is not possible to solve this problem in polynomial time unless the Exponential Time Hypothesis fails. Second, we present a strong mixed-integer programming (MIP) formulation for this problem, which is capable of exactly modeling the nonconvex and discontinuous expected reward function. Moreover, we show that this MIP formulation is ideal (i.e. the strongest possible formulation) for the revenue function of a single impression. Since it can be computationally expensive to exactly solve the MIP formulation in practice, we also study the performance of its linear programming (LP) relaxation. Though it may work well in practice, we show that, unfortunately, in the worst case the optimal objective of the LP relaxation can be $O(\text{number of samples})$ times larger than the optimal objective of the true problem. Finally, we present computational results, showcasing that the MIP formulation, along with its LP relaxation, are able to achieve superior in- and out-of-sample performance, as compared to state-of-the-art algorithms on both real and synthetic datasets. More broadly, we believe this work offers an indication of the strength of optimization methodologies like MIP to exactly model intrinsic discontinuities in machine learning problems.

[ExpandNets: Linear Over-parameterization to Train Compact Convolutional Networks](#)

- Shuxuan Guo, Jose M. Alvarez, Mathieu Salzmann
- abstract@[open-review](#): We introduce an approach to training a given compact network. To this end, we leverage over-parameterization, which typically improves both neural network optimization and generalization. Specifically, we propose to expand each linear layer of the compact network into multiple consecutive linear layers, without adding any nonlinearity. As such, the resulting expanded network, or ExpandNet, can be contracted back to the compact one algebraically at inference. In particular, we introduce two convolutional expansion strategies and demonstrate their benefits on several tasks, including image classification, object detection, and semantic segmentation. As evidenced by our experiments, our approach outperforms both training the compact network from scratch and performing knowledge distillation from a teacher. Furthermore, our linear over-parameterization empirically reduces gradient confusion during training and improves the network generalization.

[FleXOR: Trainable Fractional Quantization](#)

- Dongsoo Lee, Se Jung Kwon, Byeongwook Kim, Yongkweon Jeon, Baeseong Park, Jeongin Yun
- abstract@[open-review](#): Quantization based on the binary codes is gaining attention because each quantized bit can be directly utilized for computations without dequantization using look-up tables. Previous attempts, however, only allow for integer numbers of quantization bits, which ends up restricting the search space for compression ratio and accuracy. In this paper, we propose an encryption algorithm/architecture to compress quantized weights so as to achieve fractional numbers of bits per weight. Decryption during inference is implemented by digital XOR-gate networks added into the neural network model while XOR gates are described by utilizing $\tanh(x)$ for backward propagation to enable gradient calculations. We perform experiments using MNIST, CIFAR-10, and ImageNet to show that inserting XOR gates learns quantization/encrypted bit decisions through training and obtains high accuracy even for fractional sub 1-bit weights. As a result, our proposed method yields smaller size and higher model accuracy compared to binary neural networks.

[The Implications of Local Correlation on Learning Some Deep Functions](#)

- Eran Malach, Shai Shalev-Shwartz
- abstract@[open-review](#): It is known that learning deep neural-networks is computationally hard in the worst-case. In fact, the proofs of such hardness results show that even weakly learning deep networks is hard. In other words, no efficient algorithm can find a predictor that is slightly better than a random guess. However, we observe that on natural distributions of images, small patches of the input image are correlated to the target label, which implies that on such natural data, efficient weak learning is trivial. While in the distribution-free setting, the celebrated boosting results show that weak learning implies strong learning, in the distribution-specific setting this is not necessarily the case. We introduce a property of distributions, denoted “local correlation”, which requires that small patches of the input image and of intermediate layers of the target function are correlated to the target label. We empirically demonstrate that this property holds for the CIFAR and ImageNet data sets. The main technical results of the paper is proving that, for some classes of deep functions, weak learning implies efficient strong learning under the “local correlation” assumption.

[Learning to search efficiently for causally near-optimal treatments](#)

- Samuel Håkansson, Viktor Lindblom, Omer Gottesman, Fredrik D. Johansson
- abstract@[open-review](#): Finding an effective medical treatment often requires a search by trial and error. Making this search more efficient by minimizing the number of unnecessary trials could lower both costs and patient suffering. We formalize this problem as learning a policy for finding a near-optimal treatment in a minimum number of trials using a causal inference framework. We give a model-based dynamic programming algorithm which learns from observational data while being robust to unmeasured confounding. To reduce time complexity, we suggest a greedy algorithm which bounds the near-optimality constraint. The methods are evaluated on synthetic and real-world healthcare data and compared to model-free reinforcement learning. We find that our methods compare favorably to the model-free baseline while offering a more transparent trade-off between search time and treatment efficacy.

[A Game Theoretic Analysis of Additive Adversarial Attacks and Defenses](#)

- Ambar Pal, Rene Vidal
- abstract@[open-review](#): Research in adversarial learning follows a cat and mouse game between attackers and defenders where attacks are proposed, they are mitigated by new defenses, and subsequently new attacks are proposed that break earlier defenses, and so on. However, it has remained unclear as to whether there are conditions under which no better attacks or defenses can be proposed. In this paper, we propose a game-theoretic framework for studying attacks and defenses which exist in equilibrium. Under a locally linear decision boundary model for the underlying binary classifier, we prove that the Fast Gradient Method attack and a Randomized Smoothing defense form a Nash Equilibrium. We then show how this equilibrium defense can be approximated given finitely many samples from a data-generating distribution, and derive a generalization bound for the performance of our approximation.

[Posterior Network: Uncertainty Estimation without OOD Samples via Density-Based Pseudo-Counts](#)

- Bertrand Charpentier, Daniel ZĂ¼gner, Stephan GĂ¼nnemann
- abstract@[open-review](#): Accurate estimation of aleatoric and epistemic uncertainty is crucial to build safe and reliable systems. Traditional approaches, such as dropout and ensemble methods, estimate uncertainty by sampling probability predictions from different submodels, which leads to slow uncertainty estimation at inference time. Recent works address this drawback by directly predicting parameters of prior distributions over the probability predictions with a neural network. While this approach has demonstrated accurate uncertainty estimation, it requires defining arbitrary target parameters for in-distribution data and makes the unrealistic assumption that out-of-distribution (OOD) data is known at training time. In this work we propose the Posterior Network (PostNet), which uses Normalizing Flows to predict an individual closed-form posterior distribution over predicted probabilities for any input sample. The posterior distributions learned by PostNet accurately reflect uncertainty for in- and out-of-distribution data -- without requiring access to OOD data at training time. PostNet achieves state-of-the art results in OOD detection and in uncertainty calibration under dataset shifts.

[Recurrent Quantum Neural Networks](#)

- Johannes Bausch
- abstract@[open-review](#): Recurrent neural networks are the foundation of many sequence-to-sequence models in machine learning, such as machine translation and speech synthesis. With applied quantum computing in its infancy, there already exist quantum machine learning models such as variational quantum eigensolvers which have been used e.g. in the context of energy minimization tasks. Yet, to date, no viable recurrent quantum network has been proposed. In this work we construct the first quantum recurrent neural network (QRNN) with demonstrable performance on non-trivial tasks such as sequence learning and integer digit classification. The QRNN cell is built from parametrized quantum neurons, which, in conjunction with amplitude amplification, creates a nonlinear activation of polynomials of its inputs and hidden state, and allows the extraction of a probability distribution over predicted classes at each step.

To study the model's performance, we provide an implementation in pytorch, which allows the relatively efficient optimization of parametrized quantum circuits with tens of thousands of parameters, and which demonstrates that model does not appear to suffer from the vanishing gradient problem that plagues many existing quantum classifiers and classical RNNs. We establish a QRNN training setup by benchmarking optimization hyperparameters, and analyse suitable network topologies for simple memorisation and sequence prediction tasks from Elman's seminal paper (1990). We then proceed to evaluate the QRNN on MNIST classification, by feeding the QRNN each image pixel-by-pixel; with a network utilizing only 12 qubits we reach a test set accuracy over 98% when discriminating between the digits 0' and 1'.

[No-Regret Learning and Mixed Nash Equilibria: They Do Not Mix](#)

- Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, Thanasis Lianeas, Panayotis Mertikopoulos, Georgios Piliouras
- abstract@[open-review](#): Understanding the behavior of no-regret dynamics in general N-player games is a fundamental question in online learning and game theory. A folk result in the field states that, in finite games, the empirical frequency of play under no-regret learning converges to the gameâ€™s set of coarse correlated equilibria. By contrast, our understanding of how the day-to-day behavior of the dynamics correlates to the gameâ€™s Nash equilibria is much more limited, and only partial results are known for certain classes of games (such as zero-sum or congestion games). In this paper, we study the dynamics of follow the regularized leader (FTRL), arguably the most well-studied class of no-regret dynamics, and we establish a sweeping negative result showing that the notion of mixed Nash equilibrium is antithetical to no-regret learning. Specifically, we show that any Nash equilibrium which is not strict (in that every player has a unique best response) cannot be stable and attracting under the dynamics of FTRL. This result has significant implications for predicting the outcome of a learning process as it shows unequivocally that only strict (and hence, pure) Nash equilibria can emerge as stable limit points thereof.

[A Unifying View of Optimism in Episodic Reinforcement Learning](#)

- Gergely Neu, Ciara Pike-Burke
- abstract@[open-review](#): The principle of ``optimism in the face of uncertainty'' underpins many theoretically successful reinforcement learning algorithms. In this paper we provide a general framework for designing, analyzing and implementing such algorithms in the episodic reinforcement learning problem. This framework is built upon Lagrangian duality, and demonstrates that every model-optimistic algorithm that constructs an optimistic MDP has an equivalent representation as a value-optimistic dynamic programming algorithm. Typically, it was thought that these two classes of algorithms were distinct, with model-optimistic algorithms benefiting from a cleaner probabilistic analysis while value-optimistic algorithms are easier to implement and thus more practical. With the framework developed in this paper, we show that it is possible to get the best of both worlds by providing a class of algorithms which have a computationally efficient dynamic-programming implementation and also a simple probabilistic analysis. Besides being able to capture many existing algorithms in the tabular setting, our framework can also address large-scale problems under realizable function approximation, where it enables a simple model-based analysis of some recently proposed methods.

[Continuous Submodular Maximization: Beyond DR-Submodularity](#)

- Moran Feldman, Amin Karbasi
- abstract@[open-review](#): In this paper, we propose the first continuous optimization algorithms that achieve a constant factor approximation guarantee for the problem of monotone continuous submodular maximization subject to a linear constraint. We first prove that a simple variant of the vanilla coordinate ascent, called \COORDINATE-ASCENT+, achieves a $\$O(\frac{e-1}{2e-1}\cdot\epsilon)$ -approximation guarantee while performing $\$O(n/\epsilon)$ iterations, where the computational complexity of each iteration is roughly $\$O(n\sqrt{\epsilon}+n\log n)$ (here, n denotes the dimension of the optimization problem). We then propose \COORDINATE-ASCENT++, that achieves the tight $\$(1-1/e)\epsilon$ -approximation guarantee while performing the same number of iterations, but at a higher computational complexity of roughly $\$O(n^3\epsilon^{2.5} + n^3\log n / \epsilon^2)$ per iteration. However, the computation of each round of \COORDINATE-ASCENT++ can be easily parallelized so that the computational cost per machine scales as $\$O(n\sqrt{\epsilon}+n\log n)$.

[An Asymptotically Optimal Primal-Dual Incremental Algorithm for Contextual Linear Bandits](#)

- Andrea Tirinzoni, Matteo Pirotta, Marcello Restelli, Alessandro Lazaric

- abstract@[open-review](#): In the contextual linear bandit setting, algorithms built on the optimism principle fail to exploit the structure of the problem and have been shown to be asymptotically suboptimal. In this paper, we follow recent approaches of deriving asymptotically optimal algorithms from problem-dependent regret lower bounds and we introduce a novel algorithm improving over the state-of-the-art along multiple dimensions. We build on a reformulation of the lower bound, where context distribution and exploration policy are decoupled, and we obtain an algorithm robust to unbalanced context distributions. Then, using an incremental primal-dual approach to solve the Lagrangian relaxation of the lower bound, we obtain a scalable and computationally efficient algorithm. Finally, we remove forced exploration and build on confidence intervals of the optimization problem to encourage a minimum level of exploration that is better adapted to the problem structure. We demonstrate the asymptotic optimality of our algorithm, while providing both problem-dependent and worst-case finite-time regret guarantees. Our bounds scale with the logarithm of the number of arms, thus avoiding the linear dependence common in all related prior works. Notably, we establish minimax optimality for any learning horizon in the special case of non-contextual linear bandits. Finally, we verify that our algorithm obtains better empirical performance than state-of-the-art baselines.

[Assessing SATNet's Ability to Solve the Symbol Grounding Problem](#)

- Oscar Chang, Lampros Flokas, Hod Lipson, Michael Spranger
- abstract@[open-review](#): SATNet is an award-winning MAXSAT solver that can be used to infer logical rules and integrated as a differentiable layer in a deep neural network. It had been shown to solve Sudoku puzzles visually from examples of puzzle digit images, and was heralded as an impressive achievement towards the longstanding AI goal of combining pattern recognition with logical reasoning. In this paper, we clarify SATNet's capabilities by showing that in the absence of intermediate labels that identify individual Sudoku digit images with their logical representations, SATNet completely fails at visual Sudoku (0% test accuracy). More generally, the failure can be pinpointed to its inability to learn to assign symbols to perceptual phenomena, also known as the symbol grounding problem, which has long been thought to be a prerequisite for intelligent agents to perform real-world logical reasoning. We propose an MNIST based test as an easy instance of the symbol grounding problem that can serve as a sanity check for differentiable symbolic solvers in general. Naive applications of SATNet on this test lead to performance worse than that of models without logical reasoning capabilities. We report on the causes of SATNet's failure and how to prevent them.

[A Bayesian Nonparametrics View into Deep Representations](#)

- MichaÅ, JamroÅ¼, Marcin Kurdziel, Mateusz Opala
- abstract@[open-review](#): We investigate neural network representations from a probabilistic perspective. Specifically, we leverage Bayesian nonparametrics to construct models of neural activations in Convolutional Neural Networks (CNNs) and latent representations in Variational Autoencoders (VAEs). This allows us to formulate a tractable complexity measure for distributions of neural activations and to explore global structure of latent spaces learned by VAEs. We use this machinery to uncover how memorization and two common forms of regularization, i.e. dropout and input augmentation, influence representational complexity in CNNs. We demonstrate that networks that can exploit patterns in data learn vastly less complex representations than networks forced to memorize. We also show marked differences between effects of input augmentation and dropout, with the latter strongly depending on network width. Next, we investigate latent representations learned by standard β -VAEs and Maximum Mean Discrepancy (MMD) β -VAEs. We show that aggregated posterior in standard VAEs quickly collapses to the diagonal prior when regularization strength increases. MMD-VAEs, on the other hand, learn more complex posterior distributions, even with strong regularization. While this gives a richer sample space, MMD-VAEs do not exhibit independence of latent dimensions. Finally, we leverage our probabilistic models as an effective sampling strategy for latent codes, improving quality of samples in VAEs with rich posteriors.

[On the Similarity between the Laplace and Neural Tangent Kernels](#)

- Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, Basri Ronen
- abstract@[open-review](#): Recent theoretical work has shown that massively overparameterized neural networks are equivalent to kernel regressors that use Neural Tangent Kernels (NTKs). Experiments show that these kernel methods perform similarly to real neural networks. Here we show that NTK for fully connected networks with ReLU activation is closely related to the standard Laplace kernel. We show theoretically that for normalized data on the hypersphere both kernels have the same eigenfunctions and their eigenvalues decay polynomially at the same rate, implying that their Reproducing Kernel Hilbert Spaces (RKHS) include the same sets of functions. This means that both kernels give rise to classes of functions with the same smoothness properties. The two kernels differ for data off the hypersphere, but experiments indicate that when data is properly normalized these differences are not significant. Finally, we provide experiments on real data comparing NTK and the Laplace kernel, along with a larger class of γ -exponential kernels. We show that these perform almost identically. Our results suggest that much insight about neural networks can be obtained from analysis of the well-known Laplace kernel, which has a simple closed form.

[A causal view of compositional zero-shot recognition](#)

- Yuval Atzmon, Felix Kreuk, Uri Shalit, Gal Chechik
- abstract@[open-review](#): People easily recognize new visual categories that are new combinations of known components. This compositional generalization capacity is critical for learning in real-world domains like vision and language because the long tail of new combinations dominates the distribution. Unfortunately, learning systems struggle with compositional generalization because they often build on features that are correlated with class labels even if they are not "essential" for the class. This leads to consistent misclassification of samples from a new distribution, like new combinations of known components. Here we describe an approach for compositional generalization that builds on causal ideas. First, we describe compositional zero-shot learning from a causal perspective, and propose to view zero-shot inference as finding "which intervention caused the image?". Second, we present a causal-inspired embedding model that learns disentangled representations of elementary components of visual objects from correlated (confounded) training data. We evaluate this approach on two datasets for predicting new combinations of attribute-object pairs: A well-controlled synthesized images dataset and a real world dataset which consists of fine-grained types of shoes. We show improvements compared to strong baselines.

[HiPPO: Recurrent Memory with Optimal Polynomial Projections](#)

- Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, Christopher RÃ©
- abstract@[open-review](#): A central problem in learning from sequential data is representing cumulative history in an incremental fashion as more data is processed. We introduce a general framework (HiPPO) for the online compression of continuous signals and discrete time series by projection onto polynomial bases. Given a measure that specifies the importance of each time step in the past, HiPPO produces an optimal solution to a natural online function approximation problem. As special cases, our framework yields a short derivation of the recent Legendre Memory Unit (LMU) from first principles, and generalizes the ubiquitous gating mechanism of recurrent neural networks such as GRUs. This formal framework yields a new memory update mechanism (HiPPO-LegS) that scales through time to remember all history, avoiding priors on the timescale. HiPPO-LegS enjoys the theoretical benefits of timescale robustness, fast updates, and bounded gradients. By incorporating the memory dynamics into recurrent neural networks, HiPPO RNNs can empirically capture complex temporal dependencies. On the benchmark permuted MNIST dataset, HiPPO-LegS sets a new state-of-the-art accuracy of 98.3%. Finally, on a novel trajectory classification task testing robustness to out-of-distribution timescales and missing data, HiPPO-LegS outperforms RNN and neural ODE baselines by 25-40% accuracy.

[Auto Learning Attention](#)

- Benteng Ma, Jing Zhang, Yong Xia, Dacheng Tao

- abstract@[open-review](#): Attention modules have been demonstrated effective in strengthening the representation ability of a neural network via reweighting spatial or channel features or stacking both operations sequentially. However, designing the structures of different attention operations requires a bulk of computation and extensive expertise. In this paper, we devise an Auto Learning Attention (AutoLA) method, which is the first attempt on automatic attention design. Specifically, we define a novel attention module named high order group attention (HOGA) as a directed acyclic graph (DAG) where each group represents a node, and each edge represents an operation of heterogeneous attentions. A typical HOGA architecture can be searched automatically via the differential AutoLA method within 1 GPU day using the ResNet-20 backbone on CIFAR10. Further, the searched attention module can generalize to various backbones as a plug-and-play component and outperforms popular manually designed channel and spatial attentions for many vision tasks, including image classification on CIFAR100 and ImageNet, object detection and human keypoint detection on COCO dataset. The code will be released.

[CASTLE: Regularization via Auxiliary Causal Graph Discovery](#)

- Trent Kyono, Yao Zhang, Mihaela van der Schaar
- abstract@[open-review](#): Regularization improves generalization of supervised models to out-of-sample data. Prior works have shown that prediction in the causal direction (effect from cause) results in lower testing error than the anti-causal direction. However, existing regularization methods are agnostic of causality. We introduce Causal Structure Learning (CASTLE) regularization and propose to regularize a neural network by jointly learning the causal relationships between variables. CASTLE learns the causal directed acyclical graph (DAG) as an adjacency matrix embedded in the neural network's input layers, thereby facilitating the discovery of optimal predictors. Furthermore, CASTLE efficiently reconstructs only the features in the causal DAG that have a causal neighbor, whereas reconstruction-based regularizers suboptimally reconstruct all input features. We provide a theoretical generalization bound for our approach and conduct experiments on a plethora of synthetic and real publicly available datasets demonstrating that CASTLE consistently leads to better out-of-sample predictions as compared to other popular benchmark regularizers.

[Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect](#)

- Kaihua Tang, Jianqiang Huang, Hanwang Zhang
- abstract@[open-review](#): As the class size grows, maintaining a balanced dataset across many classes is challenging because the data are long-tailed in nature; it is even impossible when the sample-of-interest co-exists with each other in one collectable unit, e.g., multiple visual instances in one image. Therefore, long-tailed classification is the key to deep learning at scale. However, existing methods are mainly based on re-weighting/re-sampling heuristics that lack a fundamental theory. In this paper, we establish a causal inference framework, which not only unravels the whys of previous methods, but also derives a new principled solution. Specifically, our theory shows that the SGD momentum is essentially a confounder in long-tailed classification. On one hand, it has a harmful causal effect that misleads the tail prediction biased towards the head. On the other hand, its induced mediation also benefits the representation learning and head prediction. Our framework elegantly disentangles the paradoxical effects of the momentum, by pursuing the direct causal effect caused by an input sample. In particular, we use causal intervention in training, and counterfactual reasoning in inference, to remove the bad" while keep the good". We achieve new state-of-the-arts on three long-tailed visual recognition benchmarks: Long-tailed CIFAR-10/-100, ImageNet-LT for image classification and LVIS for instance segmentation.

[Explainable Voting](#)

- Dominik Peters, Ariel D. Procaccia, Alexandros Psomas, Zixin Zhou
- abstract@[open-review](#): The design of voting rules is traditionally guided by desirable axioms. Recent work shows that, surprisingly, the axiomatic approach can also support the generation of explanations for voting outcomes. However, no bounds on the size of these explanations is given; for all we know, they may be unbearably tedious. We prove, however, that outcomes of the important Borda rule can be explained using $\mathcal{O}(m^2)$ steps, where m is the number of alternatives. Our main technical result is a general lower bound that, in particular, implies that the foregoing bound is asymptotically tight. We discuss the significance of our results for AI and machine learning, including their potential to bolster an emerging paradigm of automated decision making called virtual democracy.

[Deep Archimedean Copulas](#)

- Chun Kai Ling, Fei Fang, J. Zico Kolter
- abstract@[open-review](#): A central problem in machine learning and statistics is to model joint densities of random variables from data. Copulas are joint cumulative distribution functions with uniform marginal distributions and are used to capture interdependencies in isolation from marginals. Copulas are widely used within statistics, but have not gained traction in the context of modern deep learning. In this paper, we introduce ACNet, a novel differentiable neural network architecture that enforces structural properties and enables one to learn an important class of copulas--Archimedean Copulas. Unlike Generative Adversarial Networks, Variational Autoencoders, or Normalizing Flow methods, which learn either densities or the generative process directly, ACNet learns a generator of the copula, which implicitly defines the cumulative distribution function of a joint distribution. We give a probabilistic interpretation of the network parameters of ACNet and use this to derive a simple but efficient sampling algorithm for the learned copula. Our experiments show that ACNet is able to both approximate common Archimedean Copulas and generate new copulas which may provide better fits to data.

[Re-Examining Linear Embeddings for High-Dimensional Bayesian Optimization](#)

- Ben Letham, Roberto Calandra, Akshara Rai, Eytan Bakshy
- abstract@[open-review](#): Bayesian optimization (BO) is a popular approach to optimize expensive-to-evaluate black-box functions. A significant challenge in BO is to scale to high-dimensional parameter spaces while retaining sample efficiency. A solution considered in existing literature is to embed the high-dimensional space in a lower-dimensional manifold, often via a random linear embedding. In this paper, we identify several crucial issues and misconceptions about the use of linear embeddings for BO. We study the properties of linear embeddings from the literature and show that some of the design choices in current approaches adversely impact their performance. We show empirically that properly addressing these issues significantly improves the efficacy of linear embeddings for BO on a range of problems, including learning a gait policy for robot locomotion.

[UnModNet: Learning to Unwrap a Modulo Image for High Dynamic Range Imaging](#)

- Chu Zhou, Hang Zhao, Jin Han, Chang Xu, Chao Xu, Tiejun Huang, Boxin Shi
- abstract@[open-review](#): A conventional camera often suffers from over- or under-exposure when recording a real-world scene with a very high dynamic range (HDR). In contrast, a modulo camera with a Markov random field (MRF) based unwrapping algorithm can theoretically accomplish unbounded dynamic range but shows degenerate performances when there are modulus-intensity ambiguity, strong local contrast, and color misalignment. In this paper, we reformulate the modulo image unwrapping problem into a series of binary labeling problems and propose a modulo edge-aware model, named as UnModNet, to iteratively estimate the binary rollover masks of the modulo image for unwrapping. Experimental results show that our approach can generate 12-bit HDR images from 8-bit modulo images reliably, and runs much faster than the previous MRF-based algorithm thanks to the GPU acceleration.

[Thunder: a Fast Coordinate Selection Solver for Sparse Learning](#)

- Shaogang Ren, Weijie Zhao, Ping Li
- abstract@[open-review](#): L1 regularization has been broadly employed to pursue model sparsity. Despite the non-smoothness, people have developed efficient algorithms by leveraging the sparsity and convexity of the problems. In this paper, we propose a novel active incremental approach to further improve the efficiency of the solvers. We show that our method performs well even when the existing methods fail due to the low sparseness or high solution accuracy request. Theoretical analysis and experimental results on synthetic and real-world data sets validate the advantages of the method.

[Neural Networks Fail to Learn Periodic Functions and How to Fix It](#)

- Liu Ziyin, Tilman Hartwig, Masahito Ueda
- abstract@[open-review](#): Previous literature offers limited clues on how to learn a periodic function using modern neural networks. We start with a study of the extrapolation properties of neural networks; we prove and demonstrate experimentally that the standard activation functions, such as ReLU, tanh, sigmoid, along with their variants, all fail to learn to extrapolate simple periodic functions. We hypothesize that this is due to their lack of a ``periodic'' inductive bias. As a fix of this problem, we propose a new activation, namely, $x + \sin^2(x)$, which achieves the desired periodic inductive bias to learn a periodic function while maintaining a favorable optimization property of the relu -based activations. Experimentally, we apply the proposed method to temperature and financial data prediction.

[Distribution Matching for Crowd Counting](#)

- Boyu Wang, Huidong Liu, Dimitris Samaras, Minh Hoai Nguyen
- abstract@[open-review](#): In crowd counting, each training image contains multiple people, where each person is annotated by a dot. Existing crowd counting methods need to use a Gaussian to smooth each annotated dot or to estimate the likelihood of every pixel given the annotated point. In this paper, we show that imposing Gaussians to annotations hurts generalization performance. Instead, we propose to use Distribution Matching for crowd COUNTing (DM-Count). In DM-Count, we use Optimal Transport (OT) to measure the similarity between the normalized predicted density map and the normalized ground truth density map. To stabilize OT computation, we include a Total Variation loss in our model. We show that the generalization error bound of DM-Count is tighter than that of the Gaussian smoothed methods. In terms of Mean Absolute Error, DM-Count outperforms the previous state-of-the-art methods by a large margin on two large-scale counting datasets, UCF-QNRF and NWPU, and achieves the state-of-the-art results on the ShanghaiTech and UCF-CC50 datasets. DM-Count reduced the error of the state-of-the-art published result by approximately 16%. Code is available at <https://github.com/cvlab-stonybrook/DM-Count>.

[Correspondence learning via linearly-invariant embedding](#)

- Riccardo Marin, Marie-Julie Rakotosaona, Simone Melzi, Maks Ovsjanikov
- abstract@[open-review](#): In this paper, we propose a fully differentiable pipeline for estimating accurate dense correspondences between 3D point clouds. The proposed pipeline is an extension and a generalization of the functional maps framework. However, instead of using the Laplace-Beltrami eigenfunctions as done in virtually all previous works in this domain, we demonstrate that learning the basis from data can both improve robustness and lead to better accuracy in challenging settings. We interpret the basis as a learned embedding into a higher dimensional space. Following the functional map paradigm the optimal transformation in this embedding space must be linear and we propose a separate architecture aimed at estimating the transformation by learning optimal descriptor functions. This leads to the first end-to-end trainable functional map-based correspondence approach in which both the basis and the descriptors are learned from data. Interestingly, we also observe that learning a canonical embedding leads to worse results, suggesting that leaving an extra linear degree of freedom to the embedding network gives it more robustness, thereby also shedding light onto the success of previous methods. Finally, we demonstrate that our approach achieves state-of-the-art results in challenging non-rigid 3D point cloud correspondence applications.

[Learning to Dispatch for Job Shop Scheduling via Deep Reinforcement Learning](#)

- Cong Zhang, Wen Song, Zhiguang Cao, Jie Zhang, Puay Siew Tan, Xu Chi
- abstract@[open-review](#): Priority dispatching rule (PDR) is widely used for solving real-world Job-shop scheduling problem (JSSP). However, the design of effective PDRs is a tedious task, requiring a myriad of specialized knowledge and often delivering limited performance. In this paper, we propose to automatically learn PDRs via an end-to-end deep reinforcement learning agent. We exploit the disjunctive graph representation of JSSP, and propose a Graph Neural Network based scheme to embed the states encountered during solving. The resulting policy network is size-agnostic, effectively enabling generalization on large-scale instances. Experiments show that the agent can learn high-quality PDRs from scratch with elementary raw features, and demonstrates strong performance against the best existing PDRs. The learned policies also perform well on much larger instances that are unseen in training.

[On Adaptive Attacks to Adversarial Example Defenses](#)

- Florian Tramer, Nicholas Carlini, Wieland Brendel, Aleksander Madry
- abstract@[open-review](#): Adaptive attacks have (rightfully) become the de facto standard for evaluating defenses to adversarial examples. We find, however, that typical adaptive evaluations are incomplete. We demonstrate that 13 defenses recently published at ICLR, ICML and NeurIPS---and which illustrate a diverse set of defense strategies---can be circumvented despite attempting to perform evaluations using adaptive attacks. While prior evaluation papers focused mainly on the end result---showing that a defense was ineffective---this paper focuses on laying out the methodology and the approach necessary to perform an adaptive attack. Some of our attack strategies are generalizable, but no single strategy would have been sufficient for all defenses. This underlines our key message that adaptive attacks cannot be automated and always require careful and appropriate tuning to a given defense. We hope that these analyses will serve as guidance on how to properly perform adaptive attacks against defenses to adversarial examples, and thus will allow the community to make further progress in building more robust models.

[Sinkhorn Natural Gradient for Generative Models](#)

- Zebang Shen, Zhenfu Wang, Alejandro Ribeiro, Hamed Hassani
- abstract@[open-review](#): We consider the problem of minimizing a functional over a parametric family of probability measures, where the parameterization is characterized via a push-forward structure. An important application of this problem is in training generative adversarial networks. In this regard, we propose a novel Sinkhorn Natural Gradient (SiNG) algorithm which acts as a steepest descent method on the probability space endowed with the Sinkhorn divergence. We show that the Sinkhorn information matrix (SIM), a key component of SiNG, has an explicit expression and can be evaluated accurately in complexity that scales logarithmically with respect to the desired accuracy. This is in sharp contrast to existing natural gradient methods that can only be carried out approximately. Moreover, in practical applications when only Monte-Carlo type integration is available, we design an empirical estimator for SIM and provide the stability analysis. In our experiments, we quantitatively compare SiNG with state-of-the-art SGD-type solvers on generative tasks to demonstrate its efficiency and efficacy of our method.

[Online Sinkhorn: Optimal Transport distances from sample streams](#)

- Arthur Mensch, Gabriel PeyrÃ©

- abstract@[open-review](#): Optimal Transport (OT) distances are now routinely used as loss functions in ML tasks. Yet, computing OT distances between arbitrary (i.e. not necessarily discrete) probability distributions remains an open problem. This paper introduces a new online estimator of entropy-regularized OT distances between two such arbitrary distributions. It uses streams of samples from both distributions to iteratively enrich a non-parametric representation of the transportation plan. Compared to the classic Sinkhorn algorithm, our method leverages new samples at each iteration, which enables a consistent estimation of the true regularized OT distance. We provide a theoretical analysis of the convergence of the online Sinkhorn algorithm, showing a nearly- $1/n$ asymptotic sample complexity for the iterate sequence. We validate our method on synthetic 1-d to 10-d data and on real 3-d shape data.

[Ultrahyperbolic Representation Learning](#)

- Marc Law, Jos Stam
- abstract@[open-review](#): In machine learning, data is usually represented in a (flat) Euclidean space where distances between points are along straight lines. Researchers have recently considered more exotic (non-Euclidean) Riemannian manifolds such as hyperbolic space which is well suited for tree-like data. In this paper, we propose a representation living on a pseudo-Riemannian manifold of constant nonzero curvature. It is a generalization of hyperbolic and spherical geometries where the non-degenerate metric tensor need not be positive definite. We provide the necessary learning tools in this geometry and extend gradient method optimization techniques. More specifically, we provide closed-form expressions for distances via geodesics and define a descent direction to minimize some objective function. Our novel framework is applied to graph representations.

[Locally-Adaptive Nonparametric Online Learning](#)

- Ilja Kuzborskij, NicolÃ² Cesa-Bianchi
- abstract@[open-review](#): One of the main strengths of online algorithms is their ability to adapt to arbitrary data sequences. This is especially important in nonparametric settings, where performance is measured against rich classes of comparator functions that are able to fit complex environments. Although such hard comparators and complex environments may exhibit local regularities, efficient algorithms, which can provably take advantage of these local patterns, are hardly known. We fill this gap by introducing efficient online algorithms (based on a single versatile master algorithm) each adapting to one of the following regularities: (i) local Lipschitzness of the competitor function, (ii) local metric dimension of the instance sequence, (iii) local performance of the predictor across different regions of the instance space. Extending previous approaches, we design algorithms that dynamically grow hierarchical $\hat{\mu}$ -nets on the instance space whose prunings correspond to different $\hat{\alpha}$ -locality profiles for the problem at hand. Using a technique based on tree experts, we simultaneously and efficiently compete against all such prunings, and prove regret bounds each scaling with a quantity associated with a different type of local regularity. When competing against $\hat{\alpha}$ -simple locality profiles, our technique delivers regret bounds that are significantly better than those proven using the previous approach. On the other hand, the time dependence of our bounds is not worse than that obtained by ignoring any local regularities.

[Compositional Generalization via Neural-Symbolic Stack Machines](#)

- Xinyun Chen, Chen Liang, Adams Wei Yu, Dawn Song, Denny Zhou
- abstract@[open-review](#): Despite achieving tremendous success, existing deep learning models have exposed limitations in compositional generalization, the capability to learn compositional rules and apply them to unseen cases in a systematic manner. To tackle this issue, we propose the Neural-Symbolic Stack Machine (NeSS). It contains a neural network to generate traces, which are then executed by a symbolic stack machine enhanced with sequence manipulation operations. NeSS combines the expressive power of neural sequence models with the recursion supported by the symbolic stack machine. Without training supervision on execution traces, NeSS achieves 100% generalization performance in four domains: the SCAN benchmark of language-driven navigation tasks, the task of few-shot learning of compositional instructions, the compositional machine translation benchmark, and context-free grammar parsing tasks.

[Graphon Neural Networks and the Transferability of Graph Neural Networks](#)

- Luana Ruiz, Luiz Chamon, Alejandro Ribeiro
- abstract@[open-review](#): Graph neural networks (GNNs) rely on graph convolutions to extract local features from network data. These graph convolutions combine information from adjacent nodes using coefficients that are shared across all nodes. Since these coefficients are shared and do not depend on the graph, one can envision using the same coefficients to define a GNN on another graph. This motivates analyzing the transferability of GNNs across graphs. In this paper we introduce graphon NNs as limit objects of GNNs and prove a bound on the difference between the output of a GNN and its limit graphon-NN. This bound vanishes with growing number of nodes if the graph convolutional filters are bandlimited in the graph spectral domain. This result establishes a tradeoff between discriminability and transferability of GNNs.

[Unreasonable Effectiveness of Greedy Algorithms in Multi-Armed Bandit with Many Arms](#)

- Mohsen Bayati, Nima Hamidi, Ramesh Johari, Khashayar Khosravi
- abstract@[open-review](#): We study the structure of regret-minimizing policies in the $\{\text{em many-armed}\}$ Bayesian multi-armed bandit problem: in particular, with k the number of arms and T the time horizon, we consider the case where $k \geq \sqrt{T}$. We first show that $\{\text{em subsampling}\}$ is a critical step for designing optimal policies. In particular, the standard UCB algorithm leads to sub-optimal regret bounds in the many-armed regime. However, a subsampled UCB (SS-UCB), which samples $\Theta(\sqrt{T})$ arms and executes UCB only on that subset, is rate-optimal. Despite theoretically optimal regret, even SS-UCB performs poorly due to excessive exploration of suboptimal arms. In particular, in numerical experiments SS-UCB performs worse than a simple greedy algorithm (and its subsampled version) that pulls the current empirical best arm at every time period. We show that these insights hold even in a contextual setting, using real-world data. These empirical results suggest a novel form of $\{\text{em free exploration}\}$ in the many-armed regime that benefits greedy algorithms. We theoretically study this new source of free exploration and find that it is deeply connected to the distribution of a certain tail event for the prior distribution of arm rewards. This is a fundamentally distinct phenomenon from free exploration as discussed in the recent literature on contextual bandits, where free exploration arises due to variation in contexts. We use this insight to prove that the subsampled greedy algorithm is rate-optimal for Bernoulli bandits when $k > \sqrt{T}$, and achieves sublinear regret with more general distributions. This is a case where theoretical rate optimality does not tell the whole story: when complemented by the empirical observations of our paper, the power of greedy algorithms becomes quite evident. Taken together, from a practical standpoint, our results suggest that in applications it may be preferable to use a variant of the greedy algorithm in the many-armed regime.

[Gamma-Models: Generative Temporal Difference Learning for Infinite-Horizon Prediction](#)

- Michael Janner, Igor Mordatch, Sergey Levine
- abstract@[open-review](#): We introduce the gamma-model, a predictive model of environment dynamics with an infinite, probabilistic horizon. Replacing standard single-step models with gamma-models leads to generalizations of the procedures that form the foundation of model-based control, including the model rollout and model-based value estimation. The gamma-model, trained with a generative reinterpretation of temporal difference learning, is a natural continuous analogue of the successor representation and a hybrid between model-free and model-based mechanisms. Like a value function, it contains information about the long-term future; like a standard predictive model, it is independent of task reward. We instantiate the gamma-model as both a

generative adversarial network and normalizing flow, discuss how its training reflects an inescapable tradeoff between training-time and testing-time compounding errors, and empirically investigate its utility for prediction and control.

[Deep Transformers with Latent Depth](#)

- Xian Li, Asa Cooper Stickland, Yuqing Tang, Xiang Kong
- abstract@[open-review](#): The Transformer model has achieved state-of-the-art performance in many sequence modeling tasks. However, how to leverage model capacity with large or variable depths is still an open challenge. We present a probabilistic framework to automatically learn which layer(s) to use by learning the posterior distributions of layer selection. As an extension of this framework, we propose a novel method to train one shared Transformer network for multilingual machine translation with different layer selection posteriors for each language pair. The proposed method alleviates the vanishing gradient issue and enables stable training of deep Transformers (e.g. 100 layers). We evaluate on WMT English-German machine translation and masked language modeling tasks, where our method outperforms existing approaches for training deeper Transformers. Experiments on multilingual machine translation demonstrate that this approach can effectively leverage increased model capacity and bring universal improvement for both many-to-one and one-to-many translation with diverse language pairs.

[Neural Mesh Flow: 3D Manifold Mesh Generation via Diffeomorphic Flows](#)

- Kunal Gupta, Manmohan Chandraker
- abstract@[open-review](#): Meshes are important representations of physical 3D entities in the virtual world. Applications like rendering, simulations and 3D printing require meshes to be manifold so that they can interact with the world like the real objects they represent. Prior methods generate meshes with great geometric accuracy but poor manifoldness. In this work, we propose NeuralMeshFlow (NMF) to generate two-manifold meshes for genus-0 shapes. Specifically, NMF is a shape auto-encoder consisting of several Neural Ordinary Differential Equation (NODE)(1) blocks that learn accurate mesh geometry by progressively deforming a spherical mesh. Training NMF is simpler compared to state-of-the-art methods since it does not require any explicit mesh-based regularization. Our experiments demonstrate that NMF facilitates several applications such as single-view mesh reconstruction, global shape parameterization, texture mapping, shape deformation and correspondence. Importantly, we demonstrate that manifold meshes generated using NMF are better-suited for physically-based rendering and simulation compared to prior works.

[Statistical control for spatio-temporal MEG/EEG source imaging with desparsified multi-task Lasso](#)

- Jerome-Alexis Chevalier, Joseph Salmon, Alexandre Gramfort, Bertrand Thirion
- abstract@[open-review](#): Detecting where and when brain regions activate in a cognitive task or in a given clinical condition is the promise of non-invasive techniques like magnetoencephalography (MEG) or electroencephalography (EEG). This problem, referred to as source localization, or source imaging, poses however a high-dimensional statistical inference challenge. While sparsity promoting regularizations have been proposed to address the regression problem, it remains unclear how to ensure statistical control of false detections in this setting. Moreover, MEG/EEG source imaging requires to work with spatio-temporal data and autocorrelated noise. To deal with this, we adapt the desparsified Lasso estimator ---an estimator tailored for high dimensional linear model that asymptotically follows a Gaussian distribution under sparsity and moderate feature correlation assumptions--- to temporal data corrupted with autocorrelated noise. We call it the desparsified multi-task Lasso (d-MTLasso). We combine d-MTLasso with spatially constrained clustering to reduce data dimension and with ensembling to mitigate the arbitrary choice of clustering; the resulting estimator is called ensemble of clustered desparsified multi-task Lasso (ecd-MTLasso). With respect to the current procedures, the two advantages of ecd-MTLasso are that i)it offers statistical guarantees and ii)it allows to trade spatial specificity for sensitivity, leading to a powerful adaptive method. Extensive simulations on realistic head geometries, as well as empirical results on various MEG datasets, demonstrate the high recovery performance of ecd-MTLasso and its primary practical benefit: offer a statistically principled way to threshold MEG/EEG source maps.

[A Scalable MIP-based Method for Learning Optimal Multivariate Decision Trees](#)

- Haoran Zhu, Pavankumar Murali, Dzung Phan, Lam Nguyen, Jayant Kalagnanam
- abstract@[open-review](#): Several recent publications report advances in training optimal decision trees (ODTs) using mixed-integer programs (MIPs), due to algorithmic advances in integer programming and a growing interest in addressing the inherent suboptimality of heuristic approaches such as CART. In this paper, we propose a novel MIP formulation, based on 1-norm support vector machine model, to train a binary oblique ODT for classification problems. We further present techniques, such as cutting planes, to tighten its linear relaxation, to improve run times to reach optimality. Using 36 datasets from the University of California Irvine Machine Learning Repository, we demonstrate that our training approach outperforms its counterparts from literature in terms of out-of-sample performance (around 10% improvement in mean out-of-sample testing accuracy). Towards our goal of developing a scalable framework to train multivariate ODT on large datasets, we propose a new linear programming based data selection method to choose a subset of the data, and use it to train a decision tree through our proposed MIP model. We conclude this paper with extensive numerical testing results, that showcase the generalization performance of our new MIP formulation, and the improvement in mean out-of-sample accuracy on large datasets.

[Efficient Exact Verification of Binarized Neural Networks](#)

- Kai Jia, Martin Rinard
- abstract@[open-review](#): Concerned with the reliability of neural networks, researchers have developed verification techniques to prove their robustness. Most verifiers work with real-valued networks. Unfortunately, the exact (complete and sound) verifiers face scalability challenges and provide no correctness guarantees due to floating point errors. We argue that Binarized Neural Networks (BNNs) provide comparable robustness and allow exact and significantly more efficient verification. We present a new system, EEV, for efficient and exact verification of BNNs. EEV consists of two parts: (i) a novel SAT solver that speeds up BNN verification by natively handling the reified cardinality constraints arising in BNN encodings; and (ii) strategies to train solver-friendly robust BNNs by inducing balanced layer-wise sparsity and low cardinality bounds, and adaptively cancelling the gradients. We demonstrate the effectiveness of EEV by presenting the first exact verification results for L-inf-bounded adversarial robustness of nontrivial convolutional BNNs on the MNIST and CIFAR10 datasets. Compared to exact verification of real-valued networks of the same architectures on the same tasks, EEV verifies BNNs hundreds to thousands of times faster, while delivering comparable verifiable accuracy in most cases.

[Ultra-Low Precision 4-bit Training of Deep Neural Networks](#)

- Xiao Sun, Naigang Wang, Chia-Yu Chen, Jiamin Ni, Ankur Agrawal, Xiaodong Cui, Swagath Venkataramani, Kaoutar El Maghraoui, Vijayalakshmi (Viji) Srinivasan, Kailash Gopalakrishnan
- abstract@[open-review](#): In this paper, we propose a number of novel techniques and numerical representation formats that enable, for the very first time, the precision of training systems to be aggressively scaled from 8-bits to 4-bits. To enable this advance, we explore a novel adaptive Gradient Scaling technique (Gradscale) that addresses the challenges of insufficient range and resolution in quantized gradients as well as explores the impact of quantization errors observed during model training. We theoretically analyze the role of bias in gradient quantization and propose solutions that mitigate the impact of this bias on model convergence. Finally, we examine our techniques on a spectrum of deep learning models in computer vision, speech, and NLP. In combination with previously proposed solutions for 4-bit quantization of weight and activation tensors, 4-bit training shows a non-significant loss in accuracy across application domains while enabling significant hardware acceleration (> 7X over state-of-the-art FP16 systems).

[Bridging the Gap between Sample-based and One-shot Neural Architecture Search with BONAS](#)

- Han Shi, Renjie Pi, Hang Xu, Zhengu Li, James Kwok, Tong Zhang
- abstract@[open-review](#): Neural Architecture Search (NAS) has shown great potentials in finding better neural network designs. Sample-based NAS is the most reliable approach which aims at exploring the search space and evaluating the most promising architectures. However, it is computationally very costly. As a remedy, the one-shot approach has emerged as a popular technique for accelerating NAS using weight-sharing. However, due to the weight-sharing of vastly different networks, the one-shot approach is less reliable than the sample-based approach. In this work, we propose BONAS (Bayesian Optimized Neural Architecture Search), a sample-based NAS framework which is accelerated using weight-sharing to evaluate multiple related architectures simultaneously. Specifically, we apply Graph Convolutional Network predictor as a surrogate model for Bayesian Optimization to select multiple related candidate models in each iteration. We then apply weight-sharing to train multiple candidate models simultaneously. This approach not only accelerates the traditional sample-based approach significantly, but also keeps its reliability. This is because weight-sharing among related architectures are more reliable than those in the one-shot approach. Extensive experiments are conducted to verify the effectiveness of our method over many competing algorithms.

[On Numerosity of Deep Neural Networks](#)

- Xi Zhang, Xiaolin Wu
- abstract@[open-review](#): Recently, a provocative claim was published that number sense spontaneously emerges in a deep neural network trained merely for visual object recognition. This has, if true, far reaching significance to the fields of machine learning and cognitive science alike. In this paper, we prove the above claim to be unfortunately incorrect. The statistical analysis to support the claim is flawed in that the sample set used to identify number-aware neurons is too small, compared to the huge number of neurons in the object recognition network. By this flawed analysis one could mistakenly identify number-sensing neurons in any randomly initialized deep neural networks that are not trained at all. With the above critique we ask the question what if a deep convolutional neural network is carefully trained for numerosity? Our findings are mixed. Even after being trained with number-depicting images, the deep learning approach still has difficulties to acquire the abstract concept of numbers, a cognitive task that preschoolers perform with ease. But on the other hand, we do find some encouraging evidences suggesting that deep neural networks are more robust to distribution shift for small numbers than for large numbers.

[Outlier Robust Mean Estimation with Subgaussian Rates via Stability](#)

- Ilias Diakonikolas, Daniel M. Kane, Ankit Pensia
- abstract@[open-review](#): We study the problem of outlier robust high-dimensional mean estimation under a bounded covariance assumption, and more broadly under bounded low-degree moment assumptions. We consider a standard stability condition from the recent robust statistics literature and prove that, except with exponentially small failure probability, there exists a large fraction of the inliers satisfying this condition. As a corollary, it follows that a number of recently developed algorithms for robust mean estimation, including iterative filtering and non-convex gradient descent, give optimal error estimators with (near-)subgaussian rates. Previous analyses of these algorithms gave significantly suboptimal rates. As a corollary of our approach, we obtain the first computationally efficient algorithm for outlier robust mean estimation with subgaussian rates under a bounded covariance assumption.

[Self-Supervised Relationship Probing](#)

- Jiuxiang Gu, Jason Kuen, Shafiq Joty, Jianfei Cai, Vlad Morariu, Handong Zhao, Tong Sun
- abstract@[open-review](#): Structured representations of images that model visual relationships are beneficial for many vision and vision-language applications. However, current human-annotated visual relationship datasets suffer from the long-tailed predicate distribution problem which limits the potential of visual relationship models. In this work, we introduce a self-supervised method that implicitly learns the visual relationships without relying on any ground-truth visual relationship annotations. Our method relies on 1) intra- and inter-modality encodings to respectively model relationships within each modality separately and jointly, and 2) relationship probing, which seeks to discover the graph structure within each modality. By leveraging masked language modeling, contrastive learning, and dependency tree distances for self-supervision, our method learns better object features as well as implicit visual relationships. We verify the effectiveness of our proposed method on various vision-language tasks that benefit from improved visual relationship understanding.

[Information Theoretic Counterfactual Learning from Missing-Not-At-Random Feedback](#)

- Zifeng Wang, Xi Chen, Rui Wen, Shao-Lun Huang, Ercan Kuruoglu, Yefeng Zheng
- abstract@[open-review](#): Counterfactual learning for dealing with missing-not-at-random data (MNAR) is an intriguing topic in the recommendation literature, since MNAR data are ubiquitous in modern recommender systems. Instead, missing-at-random (MAR) data, namely randomized controlled trials (RCTs), are usually required by most previous counterfactual learning methods. However, the execution of RCTs is extraordinarily expensive in practice. To circumvent the use of RCTs, we build an information theoretic counterfactual variational information bottleneck (CVIB), as an alternative for debiasing learning without RCTs. By separating the task-aware mutual information term in the original information bottleneck Lagrangian into factual and counterfactual parts, we derive a contrastive information loss and an additional output confidence penalty, which facilitates balanced learning between the factual and counterfactual domains. Empirical evaluation on real-world datasets shows that our CVIB significantly enhances both shallow and deep models, which sheds light on counterfactual learning in recommendation that goes beyond RCTs.

[Prophet Attention: Predicting Attention with Future Attention](#)

- Fenglin Liu, Xuancheng Ren, Xian Wu, Shen Ge, Wei Fan, Yuexian Zou, Xu Sun
- abstract@[open-review](#): Recently, attention based models have been used extensively in many sequence-to-sequence learning systems. Especially for image captioning, the attention based models are expected to ground correct image regions with proper generated words. However, for each time step in the decoding process, the attention based models usually use the hidden state of the current input to attend to the image regions. Under this setting, these attention models have a "deviated focus" problem that they calculate the attention weights based on previous words instead of the one to be generated, impairing the performance of both grounding and captioning. In this paper, we propose the Prophet Attention, similar to the form of self-supervision. In the training stage, this module utilizes the future information to calculate the "ideal" attention weights towards image regions. These calculated "ideal" weights are further used to regularize the "deviated" attention. In this manner, image regions are grounded with the correct words. The proposed Prophet Attention can be easily incorporated into existing image captioning models to improve their performance of both grounding and captioning. The experiments on the Flickr30k Entities and the MSCOCO datasets show that the proposed Prophet Attention consistently outperforms baselines in both automatic metrics and human evaluations. It is worth noticing that we set new state-of-the-arts on the two benchmark datasets and achieve the 1st place on the leaderboard of the online MSCOCO benchmark in terms of the default ranking score, i.e., CIDEr-c40.

[Language Models are Few-Shot Learners](#)

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu,

Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei

- abstract@[open-review](#): We demonstrate that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even becoming competitive with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks. We also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora.

Margins are Insufficient for Explaining Gradient Boosting

- Allan Grønlund, Lior Kamma, Kasper Green Larsen
- abstract@[open-review](#): Boosting is one of the most successful ideas in machine learning, achieving great practical performance with little fine-tuning. The success of boosted classifiers is most often attributed to improvements in margins. The focus on margin explanations was pioneered in the seminal work by Schapire et al. (1998) and has culminated in the $\$k\$$ th margin generalization bound by Gao and Zhou (2013), which was recently proved to be near-tight for some data distributions (Grønlund et al. 2019). In this work, we first demonstrate that the $\$k\$$ th margin bound is inadequate in explaining the performance of state-of-the-art gradient boosters. We then explain the short comings of the $\$k\$$ th margin bound and prove a stronger and more refined margin-based generalization bound that indeed succeeds in explaining the performance of modern gradient boosters. Finally, we improve upon the recent generalization lower bound by Grønlund et al. (2019).

Fourier-transform-based attribution priors improve the interpretability and stability of deep learning models for genomics

- Alex Tseng, Avanti Shrikumar, Anshul Kundaje
- abstract@[open-review](#): Deep learning models can accurately map genomic DNA sequences to associated functional molecular readouts such as protein-DNA binding data. Base-resolution importance (i.e. "attribution") scores inferred from these models can highlight predictive sequence motifs and syntax. Unfortunately, these models are prone to overfitting and are sensitive to random initializations, often resulting in noisy and irreproducible attributions that obfuscate underlying motifs. To address these shortcomings, we propose a novel attribution prior, where the Fourier transform of input-level attribution scores are computed at training-time, and high-frequency components of the Fourier spectrum are penalized. We evaluate different model architectures with and without our attribution prior, training on genome-wide binary labels or continuous molecular profiles. We show that our attribution prior significantly improves models' stability, interpretability, and performance on held-out data, especially when training data is severely limited. Our attribution prior also allows models to identify biologically meaningful sequence motifs more sensitively and precisely within individual regulatory elements. The prior is agnostic to the model architecture or predicted experimental assay, yet provides similar gains across all experiments. This work represents an important advancement in improving the reliability of deep learning models for deciphering the regulatory code of the genome.

MomentumRNN: Integrating Momentum into Recurrent Neural Networks

- Tan Nguyen, Richard Baraniuk, Andrea Bertozzi, Stanley Osher, Bao Wang
- abstract@[open-review](#): Designing deep neural networks is an art that often involves an expensive search over candidate architectures. To overcome this for recurrent neural nets (RNNs), we establish a connection between the hidden state dynamics in an RNN and gradient descent (GD). We then integrate momentum into this framework and propose a new family of RNNs, called {\\em MomentumRNNs}. We theoretically prove and numerically demonstrate that MomentumRNNs alleviate the vanishing gradient issue in training RNNs. We study the momentum long-short term memory (MomentumLSTM) and verify its advantages in convergence speed and accuracy over its LSTM counterpart across a variety of benchmarks. We also demonstrate that MomentumRNN is applicable to many types of recurrent cells, including those in the state-of-the-art orthogonal RNNs. Finally, we show that other advanced momentum-based optimization methods, such as Adam and Nesterov accelerated gradients with a restart, can be easily incorporated into the MomentumRNN framework for designing new recurrent cells with even better performance.

Marginal Utility for Planning in Continuous or Large Discrete Action Spaces

- Zaheen Ahmad, Levi Lelis, Michael Bowling
- abstract@[open-review](#): Sample-based planning is a powerful family of algorithms for generating intelligent behavior from a model of the environment. Generating good candidate actions is critical to the success of sample-based planners, particularly in continuous or large action spaces. Typically, candidate action generation exhausts the action space, uses domain knowledge, or more recently, involves learning a stochastic policy to provide such search guidance. In this paper we explore explicitly learning a candidate action generator by optimizing a novel objective, marginal utility. The marginal utility of an action generator measures the increase in value of an action over previously generated actions. We validate our approach in both curling, a challenging stochastic domain with continuous state and action spaces, and a location game with a discrete but large action space. We show that a generator trained with the marginal utility objective outperforms hand-coded schemes built on substantial domain knowledge, trained stochastic policies, and other natural objectives for generating actions for sampled-based planners.

Projected Stein Variational Gradient Descent

- Peng Chen, Omar Ghattas
- abstract@[open-review](#): The curse of dimensionality is a longstanding challenge in Bayesian inference in high dimensions. In this work, we propose a {projected Stein variational gradient descent} (pSVGD) method to overcome this challenge by exploiting the fundamental property of intrinsic low dimensionality of the data informed subspace stemming from ill-posedness of such problems. We adaptively construct the subspace using a gradient information matrix of the log-likelihood, and apply pSVGD to the much lower-dimensional coefficients of the parameter projection. The method is demonstrated to be more accurate and efficient than SVGD. It is also shown to be more scalable with respect to the number of parameters, samples, data points, and processor cores via experiments with parameters dimensions ranging from the hundreds to the tens of thousands.

Minimax Lower Bounds for Transfer Learning with Linear and One-hidden Layer Neural Networks

- Mohammadreza Mousavi Kalan, Zalan Fabian, Salman Avestimehr, Mahdi Soltanolkotabi
- abstract@[open-review](#): Transfer learning has emerged as a powerful technique for improving the performance of machine learning models on new domains where labeled training data may be scarce. In this approach a model trained for a source task, where plenty of labeled training data is available, is used as a starting point for training a model on a related target task with only few labeled training data. Despite recent empirical success of transfer learning approaches, the benefits and fundamental limits of transfer learning are poorly understood. In this paper we develop a statistical minimax framework to characterize the fundamental limits of transfer learning in the context of regression with linear and one-hidden layer neural network models. Specifically, we derive a lower-bound for the target generalization error achievable by any algorithm as a function of the number of labeled source and target data as well as appropriate notions of similarity between the source and target tasks. Our lower bound provides new insights into the benefits and limitations of transfer learning. We further corroborate our theoretical finding with various experiments.

SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks

- Fabian Fuchs, Daniel Worrall, Volker Fischer, Max Welling
- abstract@[open-review](#): We introduce the SE(3)-Transformer, a variant of the self-attention module for 3D point-clouds, which is equivariant under continuous 3D roto-translations. Equivariance is important to ensure stable and predictable performance in the presence of nuisance transformations of the data input. A positive corollary of equivariance is increased weight-tying within the model. The SE(3)-Transformer leverages the benefits of self-attention to operate on large point clouds with varying number of points, while guaranteeing SE(3)-equivariance for robustness. We evaluate our model on a toy N-body particle simulation dataset, showcasing the robustness of the predictions under rotations of the input. We further achieve competitive performance on two real-world datasets, ScanObjectNN and QM9. In all cases, our model outperforms a strong, non-equivariant attention baseline and an equivariant model without attention.

[On the equivalence of molecular graph convolution and molecular wave function with poor basis set](#)

- Masashi Tsubaki, Teruyasu Mizoguchi
- abstract@[open-review](#): In this study, we demonstrate that the linear combination of atomic orbitals (LCAO), an approximation introduced by Pauling and Lennard-Jones in the 1920s, corresponds to graph convolutional networks (GCNs) for molecules. However, GCNs involve unnecessary nonlinearity and deep architecture. We also verify that molecular GCNs are based on a poor basis function set compared with the standard one used in theoretical calculations or quantum chemical simulations. From these observations, we describe the quantum deep field (QDF), a machine learning (ML) model based on an underlying quantum physics, in particular the density functional theory (DFT). We believe that the QDF model can be easily understood because it can be regarded as a single linear layer GCN. Moreover, it uses two vanilla feedforward neural networks to learn an energy functional and a Hohenberg–Kohn map that have nonlinearities inherent in quantum physics and the DFT. For molecular energy prediction tasks, we demonstrated the viability of an “extrapolation,” in which we trained a QDF model with small molecules, tested it with large molecules, and achieved high extrapolation performance. We believe that we should move away from the competition of interpolation accuracy within benchmark datasets and evaluate ML models based on physics using an extrapolation setting; this will lead to reliable and practical applications, such as fast, large-scale molecular screening for discovering effective materials.

[The Power of Predictions in Online Control](#)

- Chenkai Yu, Guanya Shi, Soon-Jo Chung, Yisong Yue, Adam Wierman
- abstract@[open-review](#): We study the impact of predictions in online Linear Quadratic Regulator control with both stochastic and adversarial disturbances in the dynamics. In both settings, we characterize the optimal policy and derive tight bounds on the minimum cost and dynamic regret. Perhaps surprisingly, our analysis shows that the conventional greedy MPC approach is a near-optimal policy in both stochastic and adversarial settings. Specifically, for length-\$T\$ problems, MPC requires only \$O(\log T)\$ predictions to reach \$O(1)\$ dynamic regret, which matches (up to lower-order terms) our lower bound on the required prediction horizon for constant regret.

[Learning Affordance Landscapes for Interaction Exploration in 3D Environments](#)

- Tushar Nagarajan, Kristen Grauman
- abstract@[open-review](#): Embodied agents operating in human spaces must be able to master how their environment works: what objects can the agent use, and how can it use them? We introduce a reinforcement learning approach for exploration for interaction, whereby an embodied agent autonomously discovers the affordance landscape of a new unmapped 3D environment (such as an unfamiliar kitchen). Given an egocentric RGB-D camera and a high-level action space, the agent is rewarded for maximizing successful interactions while simultaneously training an image-based affordance segmentation model. The former yields a policy for acting efficiently in new environments to prepare for downstream interaction tasks, while the latter yields a convolutional neural network that maps image regions to the likelihood they permit each action, densifying the rewards for exploration. We demonstrate our idea with AI2-iTHOR. The results show agents can learn how to use new home environments intelligently and that it prepares them to rapidly address various downstream tasks like “find a knife and put it in the drawer.” Project page: <http://vision.cs.utexas.edu/projects/interaction-exploration/>

[Cooperative Multi-player Bandit Optimization](#)

- Ilai Bistritz, Nicholas Bambos
- abstract@[open-review](#): Consider a team of cooperative players that take actions in a networked-environment. At each turn, each player chooses an action and receives a reward that is an unknown function of all the players' actions. The goal of the team of players is to learn to play together the action profile that maximizes the sum of their rewards. However, players cannot observe the actions or rewards of other players, and can only get this information by communicating with their neighbors. We design a distributed learning algorithm that overcomes the informational bias players have towards maximizing the rewards of nearby players they got more information about. We assume twice continuously differentiable reward functions and constrained convex and compact action sets. Our communication graph is a random time-varying graph that follows an ergodic Markov chain. We prove that even if at every turn players take actions based only on the small random subset of the players' rewards that they know, our algorithm converges with probability 1 to the set of stationary points of (projected) gradient ascent on the sum of rewards function. Hence, if the sum of rewards is concave, then the algorithm converges with probability 1 to the optimal action profile.

[Tight First- and Second-Order Regret Bounds for Adversarial Linear Bandits](#)

- Shinji Ito, Shuichi Hirahara, Tasuku Soma, Yuichi Yoshida
- abstract@[open-review](#): We propose novel algorithms with first- and second-order regret bounds for adversarial linear bandits. These regret bounds imply that our algorithms perform well when there is an action achieving a small cumulative loss or the loss has a small variance. In addition, we need only assumptions weaker than those of existing algorithms; our algorithms work on discrete action sets as well as continuous ones without a priori knowledge about losses, and they run efficiently if a linear optimization oracle for the action set is available. These results are obtained by combining optimistic online optimization, continuous multiplicative weight update methods, and a novel technique that we refer to as distribution truncation. We also show that the regret bounds of our algorithms are tight up to polylogarithmic factors.

[Just Pick a Sign: Optimizing Deep Multitask Models with Gradient Sign Dropout](#)

- Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, Dragomir Anguelov
- abstract@[open-review](#): The vast majority of deep models use multiple gradient signals, typically corresponding to a sum of multiple loss terms, to update a shared set of trainable weights. However, these multiple updates can impede optimal training by pulling the model in conflicting directions. We present Gradient Sign Dropout (GradDrop), a probabilistic masking procedure which samples gradients at an activation layer based on their level of consistency. GradDrop is implemented as a simple deep layer that can be used in any deep net and synergizes with other gradient balancing approaches. We show that GradDrop outperforms the state-of-the-art multiloss methods within traditional multitask and transfer learning settings, and we discuss how GradDrop reveals links between optimal multiloss training and gradient stochasticity.

[A Loss Function for Generative Neural Networks Based on Watsonâ€™s Perceptual Model](#)

- Steffen Czolbe, Oswin Krause, Ingemar Cox, Christian Igel

- abstract@[open-review](#): To train Variational Autoencoders (VAEs) to generate realistic imagery requires a loss function that reflects human perception of image similarity. We propose such a loss function based on Watson's perceptual model, which computes a weighted distance in frequency space and accounts for luminance and contrast masking. We extend the model to color images, increase its robustness to translation by using the Fourier Transform, remove artifacts due to splitting the image into blocks, and make it differentiable. In experiments, VAEs trained with the new loss function generated realistic, high-quality image samples. Compared to using the Euclidean distance and the Structural Similarity Index, the images were less blurry; compared to deep neural network based losses, the new approach required less computational resources and generated images with less artifacts.

[Dynamic Fusion of Eye Movement Data and Verbal Narrations in Knowledge-rich Domains](#)

- Ervine Zheng, Qi Yu, Rui Li, Pengcheng Shi, Anne Haake
- abstract@[open-review](#): We propose to jointly analyze experts' eye movements and verbal narrations to discover important and interpretable knowledge patterns to better understand their decision-making processes. The discovered patterns can further enhance data-driven statistical models by fusing experts' domain knowledge to support complex human-machine collaborative decision-making. Our key contribution is a novel dynamic Bayesian nonparametric model that assigns latent knowledge patterns into key phases involved in complex decision-making. Each phase is characterized by a unique distribution of word topics discovered from verbal narrations and their dynamic interactions with eye movement patterns, indicating experts' special perceptual behavior within a given decision-making stage. A new split-merge-switch sampler is developed to efficiently explore the posterior state space with an improved mixing rate. Case studies on diagnostic error prediction and disease morphology categorization help demonstrate the effectiveness of the proposed model and discovered knowledge patterns.

[Scalable Multi-Agent Reinforcement Learning for Networked Systems with Average Reward](#)

- Guannan Qu, Yiheng Lin, Adam Wierman, Na Li
- abstract@[open-review](#): It has long been recognized that multi-agent reinforcement learning (MARL) faces significant scalability issues due to the fact that the size of the state and action spaces are exponentially large in the number of agents. In this paper, we identify a rich class of networked MARL problems where the model exhibits a local dependence structure that allows it to be solved in a scalable manner. Specifically, we propose a Scalable Actor-Critic (SAC) method that can learn a near optimal localized policy for optimizing the average reward with complexity scaling with the state-action space size of local neighborhoods, as opposed to the entire network. Our result centers around identifying and exploiting an exponential decay property that ensures the effect of agents on each other decays exponentially fast in their graph distance.

[Optimizing Neural Networks via Koopman Operator Theory](#)

- Akshunna S. Dogra, William Redman
- abstract@[open-review](#): Koopman operator theory, a powerful framework for discovering the underlying dynamics of nonlinear dynamical systems, was recently shown to be intimately connected with neural network training. In this work, we take the first steps in making use of this connection. As Koopman operator theory is a linear theory, a successful implementation of it in evolving network weights and biases offers the promise of accelerated training, especially in the context of deep networks, where optimization is inherently a non-convex problem. We show that Koopman operator theoretic methods allow for accurate predictions of weights and biases of feedforward, fully connected deep networks over a non-trivial range of training time. During this window, we find that our approach is >10x faster than various gradient descent based methods (e.g. Adam, Adadelta, Adagrad), in line with our complexity analysis. We end by highlighting open questions in this exciting intersection between dynamical systems and neural network theory. We highlight additional methods by which our results could be expanded to broader classes of networks and larger training intervals, which shall be the focus of future work.

[SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence](#)

- Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, Philippe Rigollet
- abstract@[open-review](#): Stein Variational Gradient Descent (SVGD), a popular sampling algorithm, is often described as the kernelized gradient flow for the Kullback-Leibler divergence in the geometry of optimal transport. We introduce a new perspective on SVGD that instead views SVGD as the kernelized gradient flow of the chi-squared divergence. Motivated by this perspective, we provide a convergence analysis of the chi-squared gradient flow. We also show that our new perspective provides better guidelines for choosing effective kernels for SVGD.

[Adversarial Robustness of Supervised Sparse Coding](#)

- Jeremias Sulam, Ramchandran Muthukumar, Raman Arora
- abstract@[open-review](#): Several recent results provide theoretical insights into the phenomena of adversarial examples. Existing results, however, are often limited due to a gap between the simplicity of the models studied and the complexity of those deployed in practice. In this work, we strike a better balance by considering a model that involves learning a representation while at the same time giving a precise generalization bound and a robustness certificate. We focus on the hypothesis class obtained by combining a sparsity-promoting encoder coupled with a linear classifier, and show an interesting interplay between the expressivity and stability of the (supervised) representation map and a notion of margin in the feature space. We bound the robust risk (to $\$ell_2$ -bounded perturbations) of hypotheses parameterized by dictionaries that achieve a mild encoder gap on training data. Furthermore, we provide a robustness certificate for end-to-end classification. We demonstrate the applicability of our analysis by computing certified accuracy on real data, and compare with other alternatives for certified robustness.

[Differentiable Meta-Learning of Bandit Policies](#)

- Craig Boutilier, Chih-wei Hsu, Branislav Kveton, Martin Mladenov, Csaba Szepesvari, Manzil Zaheer
- abstract@[open-review](#): Exploration policies in Bayesian bandits maximize the average reward over problem instances drawn from some distribution P. In this work, we learn such policies for an unknown distribution P using samples from P. Our approach is a form of meta-learning and exploits properties of P without making strong assumptions about its form. To do this, we parameterize our policies in a differentiable way and optimize them by policy gradients, an approach that is pleasantly general and easy to implement. We derive effective gradient estimators and propose novel variance reduction techniques. We also analyze and experiment with various bandit policy classes, including neural networks and a novel softmax policy. The latter has regret guarantees and is a natural starting point for our optimization. Our experiments show the versatility of our approach. We also observe that neural network policies can learn implicit biases expressed only through the sampled instances.

[Biologically Inspired Mechanisms for Adversarial Robustness](#)

- Manish Reddy Vuuyuru, Andrzej Banburski, Nishka Pant, Tomaso Poggio
- abstract@[open-review](#): A convolutional neural network strongly robust to adversarial perturbations at reasonable computational and performance cost has not yet been demonstrated. The primate visual ventral stream seems to be robust to small perturbations in visual stimuli but the underlying mechanisms that give rise to this robust perception are not understood. In this work, we investigate the role of two biologically plausible mechanisms in adversarial robustness. We demonstrate that the non-uniform sampling performed by the primate retina and the presence of multiple receptive fields with a range of

receptive field sizes at each eccentricity improve the robustness of neural networks to small adversarial perturbations. We verify that these two mechanisms do not suffer from gradient obfuscation and study their contribution to adversarial robustness through ablation studies.

Statistical-Query Lower Bounds via Functional Gradients

- Surbhi Goel, Aravind Gollakota, Adam Klivans
- abstract@[open-review](#): We give the first statistical-query lower bounds for agnostically learning any non-polynomial activation with respect to Gaussian marginals (e.g., ReLU, sigmoid, sign). For the specific problem of ReLU regression (equivalently, agnostically learning a ReLU), we show that any statistical-query algorithm with tolerance $\$n^{\{-1/\epsilon\}^b}$ must use at least $\$2^{\{n^c\}/\epsilon}$ queries for some constants $b, c > 0$, where n is the dimension and ϵ is the accuracy parameter. Our results rule out general (as opposed to correlational) SQ learning algorithms, which is unusual for real-valued learning problems. Our techniques involve a gradient boosting procedure for "amplifying" recent lower bounds due to Diakonikolas et al. (COLT 2020) and Goel et al. (ICML 2020) on the SQ dimension of functions computed by two-layer neural networks. The crucial new ingredient is the use of a nonstandard convex functional during the boosting procedure. This also yields a best-possible reduction between two commonly studied models of learning: agnostic learning and probabilistic concepts.

Near-Optimal Reinforcement Learning with Self-Play

- Yu Bai, Chi Jin, Tiancheng Yu
- abstract@[open-review](#): This paper considers the problem of designing optimal algorithms for reinforcement learning in two-player zero-sum games. We focus on self-play algorithms which learn the optimal policy by playing against itself without any direct supervision. In a tabular episodic Markov game with S states, A max-player actions and B min-player actions, the best existing algorithm for finding an approximate Nash equilibrium requires $\mathcal{O}(S^2AB)$ steps of game playing, when only highlighting the dependency on (S,A,B) . In contrast, the best existing lower bound scales as $\Omega(S(A+B))$ and has a significant gap from the upper bound. This paper closes this gap for the first time: we propose an optimistic variant of the Nash Q-learning algorithm with sample complexity $\mathcal{O}(SAB)$, and a new Nash V-learning algorithm with sample complexity $\mathcal{O}(S(A+B))$. The latter result matches the information-theoretic lower bound in all problem-dependent parameters except for a polynomial factor of the length of each episode. In addition, we present a computational hardness result for learning the best responses against a fixed opponent in Markov games---a learning objective different from finding the Nash equilibrium.

Network Diffusions via Neural Mean-Field Dynamics

- Shushan He, Hongyuan Zha, Xiaojing Ye
- abstract@[open-review](#): We propose a novel learning framework based on neural mean-field dynamics for inference and estimation problems of diffusion on networks. Our new framework is derived from the Mori-Zwanzig formalism to obtain an exact evolution of the node infection probabilities, which renders a delay differential equation with memory integral approximated by learnable time convolution operators, resulting in a highly structured and interpretable RNN. Directly using cascade data, our framework can jointly learn the structure of the diffusion network and the evolution of infection probabilities, which are cornerstone to important downstream applications such as influence maximization. Connections between parameter learning and optimal control are also established. Empirical study shows that our approach is versatile and robust to variations of the underlying diffusion network models, and significantly outperform existing approaches in accuracy and efficiency on both synthetic and real-world data.

Self-Distillation as Instance-Specific Label Smoothing

- Zhilu Zhang, Mert Sabuncu
- abstract@[open-review](#): It has been recently demonstrated that multi-generational self-distillation can improve generalization. Despite this intriguing observation, reasons for the enhancement remain poorly understood. In this paper, we first demonstrate experimentally that the improved performance of multi-generational self-distillation is in part associated with the increasing diversity in teacher predictions. With this in mind, we offer a new interpretation for teacher-student training as amortized MAP estimation, such that teacher predictions enable instance-specific regularization. Our framework allows us to theoretically relate self-distillation to label smoothing, a commonly used technique that regularizes predictive uncertainty, and suggests the importance of predictive diversity in addition to predictive uncertainty. We present experimental results using multiple datasets and neural network architectures that, overall, demonstrate the utility of predictive diversity. Finally, we propose a novel instance-specific label smoothing technique that promotes predictive diversity without the need for a separately trained teacher model. We provide an empirical evaluation of the proposed method, which, we find, often outperforms classical label smoothing.

Towards Problem-dependent Optimal Learning Rates

- Yunbei Xu, Assaf Zeevi
- abstract@[open-review](#): We study problem-dependent rates, i.e., generalization errors that scale tightly with the variance or the effective loss at the "best hypothesis." Existing uniform convergence and localization frameworks, the most widely used tools to study this problem, often fail to simultaneously provide parameter localization and optimal dependence on the sample size. As a result, existing problem-dependent rates are often rather weak when the hypothesis class is "rich" and the worst-case bound of the loss is large. In this paper we propose a new framework based on a "uniform localized convergence" principle. We provide the first (moment-penalized) estimator that achieves the optimal variance-dependent rate for general "rich" classes; we also establish improved loss-dependent rate for standard empirical risk minimization.

Cross-lingual Retrieval for Iterative Self-Supervised Training

- Chau Tran, Yuqing Tang, Xian Li, Jiatao Gu
- abstract@[open-review](#): Recent studies have demonstrated the cross-lingual alignment ability of multilingual pretrained language models. In this work, we found that the cross-lingual alignment can be further improved by training seq2seq models on sentence pairs mined using their own encoder outputs. We utilized these findings to develop a new approach --- cross-lingual retrieval for iterative self-supervised training (CRISS), where mining and training processes are applied iteratively, improving cross-lingual alignment and translation ability at the same time. Using this method, we achieved state-of-the-art unsupervised machine translation results on 9 language directions with an average improvement of 2.4 BLEU, and on the Tatoeba sentence retrieval task in the XTREME benchmark on 16 languages with an average improvement of 21.5% in absolute accuracy. Furthermore, CRISS also brings an additional 1.8 BLEU improvement on average compared to mBART, when finetuned on supervised machine translation downstream tasks.

Rethinking pooling in graph neural networks

- Diego Mesquita, Amauri Souza, Samuel Kaski
- abstract@[open-review](#): Graph pooling is a central component of a myriad of graph neural network (GNN) architectures. As an inheritance from traditional CNNs, most approaches formulate graph pooling as a cluster assignment problem, extending the idea of local patches in regular grids to graphs. Despite the wide adherence to this design choice, no work has rigorously evaluated its influence on the success of GNNs. In this paper, we build upon representative GNNs and introduce variants that challenge the need for locality-preserving representations, either using randomization or clustering on the complement graph. Strikingly, our experiments demonstrate that using these variants does not result in any decrease in performance. To understand this

phenomenon, we study the interplay between convolutional layers and the subsequent pooling ones. We show that the convolutions play a leading role in the learned representations. In contrast to the common belief, local pooling is not responsible for the success of GNNs on relevant and widely-used benchmarks.

[Pointer Graph Networks](#)

- Petar Veličković, Lars Buesing, Matthew Overlan, Razvan Pascanu, Oriol Vinyals, Charles Blundell
- abstract@[open-review](#): Graph neural networks (GNNs) are typically applied to static graphs that are assumed to be known upfront. This static input structure is often informed purely by insight of the machine learning practitioner, and might not be optimal for the actual task the GNN is solving. In absence of reliable domain expertise, one might resort to inferring the latent graph structure, which is often difficult due to the vast search space of possible graphs. Here we introduce Pointer Graph Networks (PGNs) which augment sets or graphs with additional inferred edges for improved model generalisation ability. PGNs allow each node to dynamically point to another node, followed by message passing over these pointers. The sparsity of this adaptable graph structure makes learning tractable while still being sufficiently expressive to simulate complex algorithms. Critically, the pointing mechanism is directly supervised to model long-term sequences of operations on classical data structures, incorporating useful structural inductive biases from theoretical computer science. Qualitatively, we demonstrate that PGNs can learn parallelisable variants of pointer-based data structures, namely disjoint set unions and link/cut trees. PGNs generalise out-of-distribution to 5x larger test inputs on dynamic graph connectivity tasks, outperforming unrestricted GNNs and Deep Sets.

[Gradient Regularized V-Learning for Dynamic Treatment Regimes](#)

- Yao Zhang, Mihaela van der Schaar
- abstract@[open-review](#): Deciding how to optimally treat a patient, including how to select treatments over time among the multiple available treatments, represents one of the most important issues that need to be addressed in medicine today. A dynamic treatment regime (DTR) is a sequence of treatment rules indicating how to individualize treatments for a patient based on the previously assigned treatments and the evolving covariate history. However, DTR evaluation and learning based on offline data remain challenging problems due to the bias introduced by time-varying confounders that affect treatment assignment over time; this may lead to suboptimal treatment rules being used in practice. In this paper, we introduce Gradient Regularized V-learning (GRV), a novel method for estimating the value function of a DTR. GRV regularizes the underlying outcome and propensity score models with respect to the optimality condition in semiparametric estimation theory. On the basis of this design, we construct estimators that are efficient and stable in finite samples regime. Using multiple simulation studies and one real-world medical dataset, we demonstrate that our method is superior in DTR evaluation and learning, thereby providing improved treatment options over time for patients.

[Faster Wasserstein Distance Estimation with the Sinkhorn Divergence](#)

- Léonard Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, Gabriel Peyré
- abstract@[open-review](#): The squared Wasserstein distance is a natural quantity to compare probability distributions in a non-parametric setting. This quantity is usually estimated with the plug-in estimator, defined via a discrete optimal transport problem which can be solved to $\$\\epsilon$ -accuracy by adding an entropic regularization of order $\$\\epsilon$ and using for instance Sinkhorn's algorithm. In this work, we propose instead to estimate it with the Sinkhorn divergence, which is also built on entropic regularization but includes debiasing terms. We show that, for smooth densities, this estimator has a comparable sample complexity but allows higher regularization levels, of order $\$\\epsilon^{1/2}$, which leads to improved computational complexity bounds and a strong speedup in practice. Our theoretical analysis covers the case of both randomly sampled densities and deterministic discretizations on uniform grids. We also propose and analyze an estimator based on Richardson extrapolation of the Sinkhorn divergence which enjoys improved statistical and computational efficiency guarantees, under a condition on the regularity of the approximation error, which is in particular satisfied for Gaussian densities. We finally demonstrate the efficiency of the proposed estimators with numerical experiments.

[Forethought and Hindsight in Credit Assignment](#)

- Veronica Chelu, Doina Precup, Hado P. van Hasselt
- abstract@[open-review](#): We address the problem of credit assignment in reinforcement learning and explore fundamental questions regarding the way in which an agent can best use additional computation to propagate new information, by planning with internal models of the world to improve its predictions. Particularly, we work to understand the gains and peculiarities of planning employed as forethought via forward models or as hindsight operating with backward models. We establish the relative merits, limitations and complementary properties of both planning mechanisms in carefully constructed scenarios. Further, we investigate the best use of models in planning, primarily focusing on the selection of states in which predictions should be (re)-evaluated. Lastly, we discuss the issue of model estimation and highlight a spectrum of methods that stretch from environment dynamics predictors to planner-aware models.

[Robust Recursive Partitioning for Heterogeneous Treatment Effects with Uncertainty Quantification](#)

- Hyun-Suk Lee, Yao Zhang, William Zame, Cong Shen, Jang-Won Lee, Mihaela van der Schaar
- abstract@[open-review](#): Subgroup analysis of treatment effects plays an important role in applications from medicine to public policy to recommender systems. It allows physicians (for example) to identify groups of patients for whom a given drug or treatment is likely to be effective and groups of patients for which it is not. Most of the current methods of subgroup analysis begin with a particular algorithm for estimating individualized treatment effects (ITE) and identify subgroups by maximizing the difference across subgroups of the average treatment effect in each subgroup. These approaches have several weaknesses: they rely on a particular algorithm for estimating ITE, they ignore (in)homogeneity within identified subgroups, and they do not produce good confidence estimates. This paper develops a new method for subgroup analysis, R2P, that addresses all these weaknesses. R2P uses an arbitrary, exogenously prescribed algorithm for estimating ITE and quantifies the uncertainty of the ITE estimation, using a construction that is more robust than other methods. Experiments using synthetic and semi-synthetic datasets (based on real data) demonstrate that R2P constructs partitions that are simultaneously more homogeneous within groups and more heterogeneous across groups than the partitions produced by other methods. Moreover, because R2P can employ any ITE estimator, it also produces much narrower confidence intervals with a prescribed coverage guarantee than other methods.

[Rescuing neural spike train models from bad MLE](#)

- Diego Arribas, Yuan Zhao, Il Memming Park
- abstract@[open-review](#): The standard approach to fitting an autoregressive spike train model is to maximize the likelihood for one-step prediction. This maximum likelihood estimation (MLE) often leads to models that perform poorly when generating samples recursively for more than one time step. Moreover, the generated spike trains can fail to capture important features of the data and even show diverging firing rates. To alleviate this, we propose to directly minimize the divergence between neural recorded and model generated spike trains using spike train kernels. We develop a method that stochastically optimizes the maximum mean discrepancy induced by the kernel. Experiments performed on both real and synthetic neural data validate the proposed approach, showing that it leads to well-behaving models. Using different combinations of spike train kernels, we show that we can control the trade-off between different features which is critical for dealing with model-mismatch.

Lower Bounds and Optimal Algorithms for Personalized Federated Learning

- Filip Hanzely, Slavomír Hanzely, Samuel Horváth, Peter Richtarik
- abstract@[open-review](#): In this work, we consider the optimization formulation of personalized federated learning recently introduced by Hanzely & Richtarik (2020) which was shown to give an alternative explanation to the workings of local SGD methods. Our first contribution is establishing the first lower bounds for this formulation, for both the communication complexity and the local oracle complexity. Our second contribution is the design of several optimal methods matching these lower bounds in almost all regimes. These are the first provably optimal methods for personalized federated learning. Our optimal methods include an accelerated variant of FedProx, and an accelerated variance-reduced version of FedAvg/Local SGD. We demonstrate the practical superiority of our methods through extensive numerical experiments.

Black-Box Certification with Randomized Smoothing: A Functional Optimization Based Framework

- Dinghuai Zhang, Mao Ye, Chengyue Gong, Zhanxing Zhu, Qiang Liu
- abstract@[open-review](#): Randomized classifiers have been shown to provide a promising approach for achieving certified robustness against adversarial attacks in deep learning. However, most existing methods only leverage Gaussian smoothing noise and only work for ℓ_2 perturbation. We propose a general framework of adversarial certification with non-Gaussian noise and for more general types of attacks, from a unified functional optimization perspective. Our new framework allows us to identify a key trade-off between accuracy and robustness via designing smoothing distributions, helping to design new families of non-Gaussian smoothing distributions that work more efficiently for different ℓ_p settings, including ℓ_1 , ℓ_2 and ℓ_∞ attacks. Our proposed methods achieve better certification results than previous works and provide a new perspective on randomized smoothing certification.

Deep Imitation Learning for Bimanual Robotic Manipulation

- Fan Xie, Alexander Chowdhury, M. Clara De Paolis Kaluza, Linfeng Zhao, Lawson Wong, Rose Yu
- abstract@[open-review](#): We present a deep imitation learning framework for robotic bimanual manipulation in a continuous state-action space. A core challenge is to generalize the manipulation skills to objects in different locations. We hypothesize that modeling the relational information in the environment can significantly improve generalization. To achieve this, we propose to (i) decompose the multi-modal dynamics into elemental movement primitives, (ii) parameterize each primitive using a recurrent graph neural network to capture interactions, and (iii) integrate a high-level planner that composes primitives sequentially and a low-level controller to combine primitive dynamics and inverse kinematics control. Our model is a deep, hierarchical, modular architecture. Compared to baselines, our model generalizes better and achieves higher success rates on several simulated bimanual robotic manipulation tasks. We open source the code for simulation, data, and models at: <https://github.com/Rose-STL-Lab/HDR-IL>.

Stationary Activations for Uncertainty Calibration in Deep Learning

- Lassi Meronen, Christabella Irwanto, Arno Solin
- abstract@[open-review](#): We introduce a new family of non-linear neural network activation functions that mimic the properties induced by the widely-used Matérn family of kernels in Gaussian process (GP) models. This class spans a range of locally stationary models of various degrees of mean-square differentiability. We show an explicit link to the corresponding GP models in the case that the network consists of one infinitely wide hidden layer. In the limit of infinite smoothness the Matérn family results in the RBF kernel, and in this case we recover RBF activations. Matérn activation functions result in similar appealing properties to their counterparts in GP models, and we demonstrate that the local stationarity property together with limited mean-square differentiability shows both good performance and uncertainty calibration in Bayesian deep learning tasks. In particular, local stationarity helps calibrate out-of-distribution (OOD) uncertainty. We demonstrate these properties on classification and regression benchmarks and a radar emitter classification task.

Ensemble Distillation for Robust Model Fusion in Federated Learning

- Tao Lin, Lingjing Kong, Sebastian U. Stich, Martin Jaggi
- abstract@[open-review](#): Federated Learning (FL) is a machine learning setting where many devices collaboratively train a machine learning model while keeping the training data decentralized. In most of the current training schemes the central model is refined by averaging the parameters of the server model and the updated parameters from the client side. However, directly averaging model parameters is only possible if all models have the same structure and size, which could be a restrictive constraint in many scenarios. In this work we investigate more powerful and more flexible aggregation schemes for FL. Specifically, we propose ensemble distillation for model fusion, i.e. training the central classifier through unlabeled data on the outputs of the models from the clients. This knowledge distillation technique mitigates privacy risk and cost to the same extent as the baseline FL algorithms, but allows flexible aggregation over heterogeneous client models that can differ e.g. in size, numerical precision or structure. We show in extensive empirical experiments on various CV/NLP datasets (CIFAR-10/100, ImageNet, AG News, SST2) and settings (heterogeneous models/data) that the server model can be trained much faster, requiring fewer communication rounds than any existing FL technique so far.

Falcon: Fast Spectral Inference on Encrypted Data

- Qian Lou, Wen-jie Lu, Cheng Hong, Lei Jiang
- abstract@[open-review](#): Homomorphic Encryption (HE) based secure Neural Networks(NNs) inference is one of the most promising security solutions to emerging Machine Learning as a Service (MLaaS). In the HE-based MLaaS setting, a client encrypts the sensitive data, and uploads the encrypted data to the server that directly processes the encrypted data without decryption, and returns the encrypted result to the client. The clients' data privacy is preserved since only the client has the private key. Existing HE-enabled Neural Networks (HENNs), however, suffer from heavy computational overheads. The state-of-the-art HENNs adopt ciphertext packing techniques to reduce homomorphic multiplications by packing multiple messages into one single ciphertext. Nevertheless, rotations are required in these HENNs to implement the sum of the elements within the same ciphertext. We observed that HENNs have to pay significant computing overhead on rotations, and each of rotations is ~ 10 times more expensive than homomorphic multiplications between ciphertext and plaintext. So the massive rotations have become a primary obstacle of efficient HENNs.

In this paper, we propose a fast, frequency-domain deep neural network called Falcon, for fast inferences on encrypted data. Falcon includes a fast Homomorphic Discrete Fourier Transform (HDFT) using block-circulant matrices to homomorphically support spectral operations. We also propose several efficient methods to reduce inference latency, including Homomorphic Spectral Convolution and Homomorphic Spectral Fully Connected operations by combining the batched HE and block-circulant matrices. Our experimental results show Falcon achieves the state-of-the-art inference accuracy and reduces the inference latency by $\sim 45.45\%$ over prior HENNs on MNIST and CIFAR-10.

On Power Laws in Deep Ensembles

- Ekaterina Lobacheva, Nadezhda Chirkova, Maxim Kodryan, Dmitry P. Vetrov
- abstract@[open-review](#): Ensembles of deep neural networks are known to achieve state-of-the-art performance in uncertainty estimation and lead to accuracy improvement. In this work, we focus on a classification problem and investigate the behavior of both non-calibrated and calibrated negative log-likelihood (CNLL) of a deep ensemble as a function of the ensemble size and the member network size. We indicate the conditions under which CNLL follows a power law w. r. t. ensemble size or member network size, and analyze the dynamics of the parameters of the discovered power laws. Our

important practical finding is that one large network may perform worse than an ensemble of several medium-size networks with the same total number of parameters (we call this ensemble a memory split). Using the detected power law-like dependencies, we can predict (1) the possible gain from the ensembling of networks with given structure, (2) the optimal memory split given a memory budget, based on a relatively small number of trained networks.

[Practical Quasi-Newton Methods for Training Deep Neural Networks](#)

- Donald Goldfarb, Yi Ren, Achraf Bahamou
- abstract@[open-review](#): We consider the development of practical stochastic quasi-Newton, and in particular Kronecker-factored block diagonal BFGS and L-BFGS methods, for training deep neural networks (DNNs). In DNN training, the number of variables and components of the gradient n is often of the order of tens of millions and the Hessian has n^2 elements. Consequently, computing and storing a full n times n BFGS approximation or storing a modest number of (step, change in gradient) vector pairs for use in an L-BFGS implementation is out of the question. In our proposed methods, we approximate the Hessian by a block-diagonal matrix and use the structure of the gradient and Hessian to further approximate these blocks, each of which corresponds to a layer, as the Kronecker product of two much smaller matrices. This is analogous to the approach in KFAC , which computes a Kronecker-factored block diagonal approximation to the Fisher matrix in a stochastic natural gradient method. Because the indefinite and highly variable nature of the Hessian in a DNN, we also propose a new damping approach to keep the upper as well as the lower bounds of the BFGS and L-BFGS approximations bounded. In tests on autoencoder feed-forward network models with either nine or thirteen layers applied to three datasets, our methods outperformed or performed comparably to KFAC and state-of-the-art first-order stochastic methods.

[Approximation Based Variance Reduction for Reparameterization Gradients](#)

- Tomas Geffner, Justin Domke
- abstract@[open-review](#): Flexible variational distributions improve variational inference but are harder to optimize. In this work we present a control variate that is applicable for any reparameterizable distribution with known mean and covariance, e.g. Gaussians with any covariance structure. The control variate is based on a quadratic approximation of the model, and its parameters are set using a double-descent scheme. We empirically show that this control variate leads to large improvements in gradient variance and optimization convergence for inference with non-factorized variational distributions.

[Inference Stage Optimization for Cross-scenario 3D Human Pose Estimation](#)

- Jianfeng Zhang, Xuecheng Nie, Jiashi Feng
- abstract@[open-review](#): Existing 3D human pose estimation models suffer performance drop when applying to new scenarios with unseen poses due to their limited generalizability. In this work, we propose a novel framework, Inference Stage Optimization (ISO), for improving the generalizability of 3D pose models when source and target data come from different pose distributions. Our main insight is that the target data, even though not labeled, carry valuable priors about their underlying distribution. To exploit such information, the proposed ISO performs geometry-aware self-supervised learning (SSL) on each single target instance and updates the 3D pose model before making prediction. In this way, the model can mine distributional knowledge about the target scenario and quickly adapt to it with enhanced generalization performance. In addition, to handle sequential target data, we propose an online mode for implementing our ISO framework via streaming the SSL, which substantially enhances its effectiveness. We systematically analyze why and how our ISO framework works on diverse benchmarks under cross-scenario setup. Remarkably, it yields new state-of-the-art of 83.6% 3D PCK on MPI-INF-3DHP, improving upon the previous best result by 9.7%.

[Consistent feature selection for analytic deep neural networks](#)

- Vu C. Dinh, Lam S. Ho
- abstract@[open-review](#): One of the most important steps toward interpretability and explainability of neural network models is feature selection, which aims to identify the subset of relevant features. Theoretical results in the field have mostly focused on the prediction aspect of the problem with virtually no work on feature selection consistency for deep neural networks due to the model's severe nonlinearity and unidentifiability. This lack of theoretical foundation casts doubt on the applicability of deep learning to contexts where correct interpretations of the features play a central role. In this work, we investigate the problem of feature selection for analytic deep networks. We prove that for a wide class of networks, including deep feed-forward neural networks, convolutional neural networks and a major sub-class of residual neural networks, the Adaptive Group Lasso selection procedure with Group Lasso as the base estimator is selection-consistent. The work provides further evidence that Group Lasso might be inefficient for feature selection with neural networks and advocates the use of Adaptive Group Lasso over the popular Group Lasso.

[Glance and Focus: a Dynamic Approach to Reducing Spatial Redundancy in Image Classification](#)

- Yulin Wang, Kangchen Lv, Rui Huang, Shiji Song, Le Yang, Gao Huang
- abstract@[open-review](#): The accuracy of deep convolutional neural networks (CNNs) generally improves when fueled with high resolution images. However, this often comes at a high computational cost and high memory footprint. Inspired by the fact that not all regions in an image are task-relevant, we propose a novel framework that performs efficient image classification by processing a sequence of relatively small inputs, which are strategically selected from the original image with reinforcement learning. Such a dynamic decision process naturally facilitates adaptive inference at test time, i.e., it can be terminated once the model is sufficiently confident about its prediction and thus avoids further redundant computation. Notably, our framework is general and flexible as it is compatible with most of the state-of-the-art light-weighted CNNs (such as MobileNets, EfficientNets and RegNets), which can be conveniently deployed as the backbone feature extractor. Experiments on ImageNet show that our method consistently improves the computational efficiency of a wide variety of deep models. For example, it further reduces the average latency of the highly efficient MobileNet-V3 on an iPhone XS Max by 20% without sacrificing accuracy. Code and pre-trained models are available at <https://github.com/blackfeather-wang/GFNet-Pytorch>.

[Information Maximization for Few-Shot Learning](#)

- Malik Boudiaf, Imtiaz Ziko, JÃ©rÃ©me Rony, Jose Dolz, Pablo Piantanida, Ismail Ben Ayed
- abstract@[open-review](#): We introduce Transductive Infomation Maximization (TIM) for few-shot learning. Our method maximizes the mutual information between the query features and their label predictions for a given few-shot task, in conjunction with a supervision loss based on the support set. Furthermore, we propose a new alternating-direction solver for our mutual-information loss, which substantially speeds up transductive inference convergence over gradient-based optimization, while yielding similar accuracy. TIM inference is modular: it can be used on top of any base-training feature extractor. Following standard transductive few-shot settings, our comprehensive experiments demonstrate that TIM outperforms state-of-the-art methods significantly across various datasets and networks, while used on top of a fixed feature extractor trained with simple cross-entropy on the base classes, without resorting to complex meta-learning schemes. It consistently brings between 2% and 5% improvement in accuracy over the best performing method, not only on all the well-established few-shot benchmarks but also on more challenging scenarios, with domain shifts and larger numbers of classes.

[Inverse Reinforcement Learning from a Gradient-based Learner](#)

- Giorgia Ramponi, Gianluca Drappo, Marcello Restelli

- abstract@[open-review](#): Inverse Reinforcement Learning addresses the problem of inferring an expert's reward function from demonstrations. However, in many applications, we not only have access to the expert's near-optimal behaviour, but we also observe part of her learning process. In this paper, we propose a new algorithm for this setting, in which the goal is to recover the reward function being optimized by an agent, given a sequence of policies produced during learning. Our approach is based on the assumption that the observed agent is updating her policy parameters along the gradient direction. Then we extend our method to deal with the more realistic scenario where we only have access to a dataset of learning trajectories. For both settings, we provide theoretical insights into our algorithms' performance. Finally, we evaluate the approach in a simulated GridWorld environment and on the MuJoCo environments, comparing it with the state-of-the-art baseline.

[Bayesian Multi-type Mean Field Multi-agent Imitation Learning](#)

- Fan Yang, Alina Vereshchaka, Changyou Chen, Wen Dong
- abstract@[open-review](#): Multi-agent Imitation learning (MAIL) refers to the problem that agents learn to perform a task interactively in a multi-agent system through observing and mimicking expert demonstrations, without any knowledge of a reward function from the environment. MAIL has received a lot of attention due to promising results achieved on synthesized tasks, with the potential to be applied to complex real-world multi-agent tasks. Key challenges for MAIL include sample efficiency and scalability. In this paper, we proposed Bayesian multi-type mean field multi-agent imitation learning (BM3IL). Our method improves sample efficiency through establishing a Bayesian formulation for MAIL, and enhances scalability through introducing a new multi-type mean field approximation. We demonstrate the performance of our algorithm through benchmarking with three state-of-the-art multi-agent imitation learning algorithms on several tasks, including solving a multi-agent traffic optimization problem in a real-world transportation network. Experimental results indicate that our algorithm significantly outperforms all other algorithms in all scenarios.

[Bayesian Robust Optimization for Imitation Learning](#)

- Daniel Brown, Scott Niekum, Marek Petrik
- abstract@[open-review](#): One of the main challenges in imitation learning is determining what action an agent should take when outside the state distribution of the demonstrations. Inverse reinforcement learning (IRL) can enable generalization to new states by learning a parameterized reward function, but these approaches still face uncertainty over the true reward function and corresponding optimal policy. Existing safe imitation learning approaches based on IRL deal with this uncertainty using a maxmin framework that optimizes a policy under the assumption of an adversarial reward function, whereas risk-neutral IRL approaches either optimize a policy for the mean or MAP reward function. While completely ignoring risk can lead to overly aggressive and unsafe policies, optimizing in a fully adversarial sense is also problematic as it can lead to overly conservative policies that perform poorly in practice. To provide a bridge between these two extremes, we propose Bayesian Robust Optimization for Imitation Learning (BROIL). BROIL leverages Bayesian reward function inference and a user specific risk tolerance to efficiently optimize a robust policy that balances expected return and conditional value at risk. Our empirical results show that BROIL provides a natural way to interpolate between return-maximizing and risk-minimizing behaviors and outperforms existing risk-sensitive and risk-neutral inverse reinforcement learning algorithms.

[Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance](#)

- Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, Yaron Lipman
- abstract@[open-review](#): In this work we address the challenging problem of multiview 3D surface reconstruction. We introduce a neural network architecture that simultaneously learns the unknown geometry, camera parameters, and a neural renderer that approximates the light reflected from the surface towards the camera. The geometry is represented as a zero level-set of a neural network, while the neural renderer, derived from the rendering equation, is capable of (implicitly) modeling a wide set of lighting conditions and materials. We trained our network on real world 2D images of objects with different material properties, lighting conditions, and noisy camera initializations from the DTU MVS dataset. We found our model to produce state of the art 3D surface reconstructions with high fidelity, resolution and detail.

[Riemannian Continuous Normalizing Flows](#)

- Emile Mathieu, Maximilian Nickel
- abstract@[open-review](#): Normalizing flows have shown great promise for modelling flexible probability distributions in a computationally tractable way. However, whilst data is often naturally described on Riemannian manifolds such as spheres, torii, and hyperbolic spaces, most normalizing flows implicitly assume a flat geometry, making them either misspecified or ill-suited in these situations. To overcome this problem, we introduce Riemannian continuous normalizing flows, a model which admits the parametrization of flexible probability measures on smooth manifolds by defining flows as the solution to ordinary differential equations. We show that this approach can lead to substantial improvements on both synthetic and real-world data when compared to standard flows or previously introduced projected flows.

[Attention-Gated Brain Propagation: How the brain can implement reward-based error backpropagation](#)

- Isabella Pozzi, Sander Bohte, Pieter Roelfsema
- abstract@[open-review](#): Much recent work has focused on biologically plausible variants of supervised learning algorithms. However, there is no teacher in the motor cortex that instructs the motor neurons and learning in the brain depends on reward and punishment. We demonstrate a biologically plausible reinforcement learning scheme for deep networks with an arbitrary number of layers. The network chooses an action by selecting a unit in the output layer and uses feedback connections to assign credit to the units in successively lower layers that are responsible for this action. After the choice, the network receives reinforcement and there is no teacher correcting the errors. We show how the new learning scheme "Attention-Gated Brain Propagation (BrainProp)" is mathematically equivalent to error backpropagation, for one output unit at a time. We demonstrate successful learning of deep fully connected, convolutional and locally connected networks on classical and hard image-classification benchmarks; MNIST, CIFAR10, CIFAR100 and Tiny ImageNet. BrainProp achieves an accuracy that is equivalent to that of standard error-backpropagation, and better than state-of-the-art biologically inspired learning schemes. The trial-and-error nature of learning is associated with limited additional training time so that BrainProp is a factor of 1-3.5 times slower. Our results thereby provide new insights into how deep learning may be implemented in the brain.

[Asymptotic Guarantees for Generative Modeling Based on the Smooth Wasserstein Distance](#)

- Ziv Goldfeld, Kristjan Greenewald, Kengo Kato
- abstract@[open-review](#): Minimum distance estimation (MDE) gained recent attention as a formulation of (implicit) generative modeling. It considers minimizing, over model parameters, a statistical distance between the empirical data distribution and the model. This formulation lends itself well to theoretical analysis, but typical results are hindered by the curse of dimensionality. To overcome this and devise a scalable finite-sample statistical MDE theory, we adopt the framework of smooth 1-Wasserstein distance (SWD) $\mathbb{W}_1(\sigma)$. The SWD was recently shown to preserve the metric and topological structure of classic Wasserstein distances, while enjoying dimension-free empirical convergence rates. In this work, we conduct a thorough statistical study of the minimum smooth Wasserstein estimators (MSWEs), first proving the estimator's measurability and asymptotic consistency. We then characterize the limit distribution of the optimal model parameters and their associated minimal SWD. These results imply an $O(n^{-1/2})$ generalization bound for generative modeling based on MSWE, which holds in arbitrary dimension. Our main technical tool is a novel high-dimensional limit distribution result for empirical $\mathbb{W}_1(\sigma)$. The characterization of a nondegenerate limit stands in sharp contrast with the classic empirical 1-Wasserstein distance, for which a similar result is known only in the one-dimensional case. The validity of our theory is supported by empirical results, posing the SWD as a potent tool for learning and inference in high dimensions.

[Online Robust Regression via SGD on the l1 loss](#)

- Scott Pesme, Nicolas Flammarion
- abstract@[open-review](#): We consider the robust linear regression problem in the online setting where we have access to the data in a streaming manner, one data point after the other. More specifically, for a true parameter θ^* , we consider the corrupted Gaussian linear model $y = \theta^* + \epsilon + b$ where the adversarial noise b can take any value with probability η and equals zero otherwise. We consider this adversary to be oblivious (i.e., b independent of the data) since this is the only contamination model under which consistency is possible. Current algorithms rely on having the whole data at hand in order to identify and remove the outliers. In contrast, we show in this work that stochastic gradient descent on the l1 loss converges to the true parameter vector at a $\tilde{O}(1/(1-\eta)^2 n)$ rate which is independent of the values of the contaminated measurements. Our proof relies on the elegant smoothing of the l1 loss by the Gaussian data and a classical non-asymptotic analysis of Polyak-Ruppert averaged SGD. In addition, we provide experimental evidence of the efficiency of this simple and highly scalable algorithm.

[PRANK: motion Prediction based on RANKing](#)

- Yuriy Biktairov, Maxim Stebelev, Irina Rudenko, Oleh Shliazhko, Boris Yangel
- abstract@[open-review](#): Predicting the motion of agents such as pedestrians or human-driven vehicles is one of the most critical problems in the autonomous driving domain. The overall safety of driving and the comfort of a passenger directly depend on its successful solution. The motion prediction problem also remains one of the most challenging problems in autonomous driving engineering, mainly due to high variance of the possible agent's future behavior given a situation. The two phenomena responsible for the said variance are the multimodality caused by the uncertainty of the agent's intent (e.g., turn right or move forward) and uncertainty in the realization of a given intent (e.g., which lane to turn into). To be useful within a real-time autonomous driving pipeline, a motion prediction system must provide efficient ways to describe and quantify this uncertainty, such as computing posterior modes and their probabilities or estimating density at the point corresponding to a given trajectory. It also should not put substantial density on physically impossible trajectories, as they can confuse the system processing the predictions. In this paper, we introduce the PRANK method, which satisfies these requirements. PRANK takes rasterized bird-eye images of agent's surroundings as an input and extracts features of the scene with a convolutional neural network. It then produces the conditional distribution of agent's trajectories plausible in the given scene. The key contribution of PRANK is a way to represent that distribution using nearest-neighbor methods in latent trajectory space, which allows for efficient inference in real time. We evaluate PRANK on the in-house and Argoverse datasets, where it shows competitive results.

[Fighting Copycat Agents in Behavioral Cloning from Observation Histories](#)

- Chuan Wen, Jierui Lin, Trevor Darrell, Dinesh Jayaraman, Yang Gao
- abstract@[open-review](#): Imitation learning trains policies to map from input observations to the actions that an expert would choose. In this setting, distribution shift frequently exacerbates the effect of misattributing expert actions to nuisance correlates among the observed variables. We observe that a common instance of this causal confusion occurs in partially observed settings when expert actions are strongly correlated over time: the imitator learns to cheat by predicting the expert's previous action, rather than the next action. To combat this "copycat problem", we propose an adversarial approach to learn a feature representation that removes excess information about the previous expert action nuisance correlate, while retaining the information necessary to predict the next action. In our experiments, our approach improves performance significantly across a variety of partially observed imitation learning tasks.

[Tight Nonparametric Convergence Rates for Stochastic Gradient Descent under the Noiseless Linear Model](#)

- Raphaël Berthier, Francis Bach, Pierre Gaillard
- abstract@[open-review](#): In the context of statistical supervised learning, the noiseless linear model assumes that there exists a deterministic linear relation $Y = \langle \theta, \Phi(U) \rangle$ between the random output Y and the random feature vector $\Phi(U)$, a potentially non-linear transformation of the inputs U . We analyze the convergence of single-pass, fixed step-size stochastic gradient descent on the least-square risk under this model. The convergence of the iterates to the optimum θ^* and the decay of the generalization error follow polynomial convergence rates with exponents that both depend on the regularities of the optimum θ^* and of the feature vectors $\Phi(U)$. We interpret our result in the reproducing kernel Hilbert space framework. As a special case, we analyze an online algorithm for estimating a real function on the unit hypercube from the noiseless observation of its value at randomly sampled points; the convergence depends on the Sobolev smoothness of the function and of a chosen kernel. Finally, we apply our analysis beyond the supervised learning setting to obtain convergence rates for the averaging process (a.k.a. gossip algorithm) on a graph depending on its spectral dimension.

[Structured Prediction for Conditional Meta-Learning](#)

- Ruohan Wang, Yiannis Demiris, Carlo Ciliberto
- abstract@[open-review](#): The goal of optimization-based meta-learning is to find a single initialization shared across a distribution of tasks to speed up the process of learning new tasks. Conditional meta-learning seeks task-specific initialization to better capture complex task distributions and improve performance. However, many existing conditional methods are difficult to generalize and lack theoretical guarantees. In this work, we propose a new perspective on conditional meta-learning via structured prediction. We derive task-adaptive structured meta-learning (TASML), a principled framework that yields task-specific objective functions by weighing meta-training data on target tasks. Our non-parametric approach is model-agnostic and can be combined with existing meta-learning methods to achieve conditioning. Empirically, we show that TASML improves the performance of existing meta-learning models, and outperforms the state-of-the-art on benchmark datasets.

[Optimal Lottery Tickets via Subset Sum: Logarithmic Over-Parameterization is Sufficient](#)

- Ankit Pensia, Shashank Rajput, Alliot Nagle, Harit Vishwakarma, Dimitris Papailiopoulos
- abstract@[open-review](#): The strong lottery ticket hypothesis (LTH) postulates that one can approximate any target neural network by only pruning the weights of a sufficiently over-parameterized random network. A recent work by Malach et al. [MYSS20] establishes the first theoretical analysis for the strong LTH: one can provably approximate a neural network of width d and depth l , by pruning a random one that is a factor $O(d^4 l^2)$ wider and twice as deep. This polynomial over-parameterization requirement is at odds with recent experimental research that achieves good approximation with networks that are a small factor wider than the target. In this work, we close the gap and offer an exponential improvement to the over-parameterization requirement for the existence of lottery tickets. We show that any target network of width d and depth l can be approximated by pruning a random network that is a factor $O(\log(dl))$ wider and twice as deep. Our analysis heavily relies on connecting pruning random ReLU networks to random instances of the Subset Sum problem. We then show that this logarithmic over-parameterization is essentially optimal for constant depth networks. Finally, we verify several of our theoretical insights with experiments.

[The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes](#)

- Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, Davide Testuggine
- abstract@[open-review](#): This work proposes a new challenge set for multimodal classification, focusing on detecting hate speech in multimodal memes. It is constructed such that unimodal models struggle and only multimodal models can succeed: difficult examples (âœbenign confoundersâœ) are added

to the dataset to make it hard to rely on unimodal signals. The task requires subtle reasoning, yet is straightforward to evaluate as a binary classification problem. We provide baseline performance numbers for unimodal models, as well as for multimodal models with various degrees of sophistication. We find that state-of-the-art methods perform poorly compared to humans, illustrating the difficulty of the task and highlighting the challenge that this important problem poses to the community.

[Stochasticity of Deterministic Gradient Descent: Large Learning Rate for Multiscale Objective Function](#)

- Lingkai Kong, Molei Tao
- abstract@[open-review](#): This article suggests that deterministic Gradient Descent, which does not use any stochastic gradient approximation, can still exhibit stochastic behaviors. In particular, it shows that if the objective function exhibit multiscale behaviors, then in a large learning rate regime which only resolves the macroscopic but not the microscopic details of the objective, the deterministic GD dynamics can become chaotic and convergent not to a local minimizer but to a statistical distribution. In this sense, deterministic GD resembles stochastic GD even though no stochasticity is injected. A sufficient condition is also established for approximating this long-time statistical limit by a rescaled Gibbs distribution, which for example allows escapes from local minima to be quantified. Both theoretical and numerical demonstrations are provided, and the theoretical part relies on the construction of a stochastic map that uses bounded noise (as opposed to Gaussian noise).

[Identifying Learning Rules From Neural Network Observables](#)

- Aran Nayebi, Sanjana Srivastava, Surya Ganguli, Daniel L. Yamins
- abstract@[open-review](#): The brain modifies its synaptic strengths during learning in order to better adapt to its environment. However, the underlying plasticity rules that govern learning are unknown. Many proposals have been suggested, including Hebbian mechanisms, explicit error backpropagation, and a variety of alternatives. It is an open question as to what specific experimental measurements would need to be made to determine whether any given learning rule is operative in a real biological system. In this work, we take a "virtual experimental" approach to this problem. Simulating idealized neuroscience experiments with artificial neural networks, we generate a large-scale dataset of learning trajectories of aggregate statistics measured in a variety of neural network architectures, loss functions, learning rule hyperparameters, and parameter initializations. We then take a discriminative approach, training linear and simple non-linear classifiers to identify learning rules from features based on these observables. We show that different classes of learning rules can be separated solely on the basis of aggregate statistics of the weights, activations, or instantaneous layer-wise activity changes, and that these results generalize to limited access to the trajectory and held-out architectures and learning curricula. We identify the statistics of each observable that are most relevant for rule identification, finding that statistics from network activities across training are more robust to unit undersampling and measurement noise than those obtained from the synaptic strengths. Our results suggest that activation patterns, available from electrophysiological recordings of post-synaptic activities on the order of several hundred units, frequently measured at wider intervals over the course of learning, may provide a good basis on which to identify learning rules.

[Optimal Approximation - Smoothness Tradeoffs for Soft-Max Functions](#)

- Alessandro Epasto, Mohammad Mahdian, Vahab Mirrokni, Emmanouil Zampetakis
- abstract@[open-review](#): A soft-max function has two main efficiency measures: (1) approximation - which corresponds to how well it approximates the maximum function, (2) smoothness - which shows how sensitive it is to changes of its input. Our goal is to identify the optimal approximation-smoothness tradeoffs for different measures of approximation and smoothness. This leads to novel soft-max functions, each of which is optimal for a different application. The most commonly used soft-max function, called exponential mechanism, has optimal tradeoff between approximation measured in terms of expected additive approximation and smoothness measured with respect to Renyi Divergence. We introduce a soft-max function, called piece-wise linear soft-max, with optimal tradeoff between approximation, measured in terms of worst-case additive approximation and smoothness, measured with respect to l_q -norm. The worst-case approximation guarantee of the piece-wise linear mechanism enforces sparsity in the output of our soft-max function, a property that is known to be important in Machine Learning applications Martins et al. 16, Laha et al. 18 and is not satisfied by the exponential mechanism. Moreover, the l_q -smoothness is suitable for applications in Mechanism Design and Game Theory where the piece-wise linear mechanism outperforms the exponential mechanism. Finally, we investigate another soft-max function, called power mechanism, with optimal tradeoff between expected multiplicative approximation and smoothness with

respect to the Renyi Divergence, which provides improved theoretical and practical results in differentially private submodular optimization.

[Weakly-Supervised Reinforcement Learning for Controllable Behavior](#)

- Lisa Lee, Ben Eysenbach, Russ R. Salakhutdinov, Shixiang (Shane) Gu, Chelsea Finn
- abstract@[open-review](#): Reinforcement learning (RL) is a powerful framework for learning to take actions to solve tasks. However, in many settings, an agent must winnow down the inconceivably large space of all possible tasks to the single task that it is currently being asked to solve. Can we instead constrain the space of tasks to those that are semantically meaningful? In this work, we introduce a framework for using weak supervision to automatically disentangle this semantically meaningful subspace of tasks from the enormous space of nonsensical "chaff" tasks. We show that this learned subspace enables efficient exploration and provides a representation that captures distance between states. On a variety of challenging, vision-based continuous control problems, our approach leads to substantial performance gains, particularly as the complexity of the environment grows.

[Improving Policy-Constrained Kidney Exchange via Pre-Screening](#)

- Duncan McElfresh, Michael Curry, Tuomas Sandholm, John Dickerson
- abstract@[open-review](#): In barter exchanges, participants swap goods with one another without exchanging money; these exchanges are often facilitated by a central clearinghouse, with the goal of maximizing the aggregate quality (or number) of swaps. Barter exchanges are subject to many forms of uncertainty--in participant preferences, the feasibility and quality of various swaps, and so on. Our work is motivated by kidney exchange, a real-world barter market in which patients in need of a kidney transplant swap their willing living donors, in order to find a better match. Modern exchanges include 2- and 3-way swaps, making the kidney exchange clearing problem NP-hard. Planned transplants often fail for a variety of reasons--if the donor organ is rejected by the recipient's medical team, or if the donor and recipient are found to be medically incompatible. Due to 2- and 3-way swaps, failed transplants can ``cascade'' through an exchange; one US-based exchange estimated that about 85% of planned transplants failed in 2019. Many optimization-based approaches have been designed to avoid these failures; however most exchanges cannot implement these methods, due to legal and policy constraints. Instead, we consider a setting where exchanges can query the preferences of certain donors and recipients--asking whether they would accept a particular transplant. We characterize this as a two-stage decision problem, in which the exchange program (a) queries a small number of transplants before committing to a matching, and (b) constructs a matching according to fixed policy. We show that selecting these edges is a challenging combinatorial problem, which is non-monotonic and non-submodular, in addition to being NP-hard. We propose both a greedy heuristic and a Monte Carlo tree search, which outperforms previous approaches, using experiments on both synthetic data and real kidney exchange data from the United Network for Organ Sharing.

[Learning abstract structure for drawing by efficient motor program induction](#)

- Lucas Tian, Kevin Ellis, Marta Kryven, Josh Tenenbaum

- abstract@[open-review](#): Humans flexibly solve new problems that differ from those previously practiced. This ability to flexibly generalize is supported by learned concepts that represent useful structure common across different problems. Here we develop a naturalistic drawing task to study how humans rapidly acquire structured prior knowledge. The task requires drawing visual figures that share underlying structure, based on a set of composable geometric rules and simple objects. We show that people spontaneously learn abstract drawing procedures that support generalization, and propose a model of how learners can discover these reusable drawing procedures. Trained in the same setting as humans, and constrained to produce efficient motor actions, this model discovers new drawing program subroutines that generalize to test figures and resemble learned features of human behavior. These results suggest that two principles guiding motor program induction in the model - abstraction (programs can reflect high-level structure that ignores figure-specific details) and compositionality (new programs are discovered by recombining previously learned programs) - are key for explaining how humans learn structured internal representations that guide flexible reasoning and learning.

[Why Do Deep Residual Networks Generalize Better than Deep Feedforward Networks? --- A Neural Tangent Kernel Perspective](#)

- Kaixuan Huang, Yuqing Wang, Molei Tao, Tuo Zhao
- abstract@[open-review](#): Deep residual networks (ResNets) have demonstrated better generalization performance than deep feedforward networks (FFNets). However, the theory behind such a phenomenon is still largely unknown. This paper studies this fundamental problem in deep learning from a so-called ``neural tangent kernel'' perspective. Specifically, we first show that under proper conditions, as the width goes to infinity, training deep ResNets can be viewed as learning reproducing kernel functions with some kernel function. We then compare the kernel of deep ResNets with that of deep FFNets and discover that the class of functions induced by the kernel of FFNets is asymptotically not learnable, as the depth goes to infinity. In contrast, the class of functions induced by the kernel of ResNets does not exhibit such degeneracy. Our discovery partially justifies the advantages of deep ResNets over deep FFNets in generalization abilities. Numerical results are provided to support our claim.

[Dual Instrumental Variable Regression](#)

- Krikamol Muandet, Arash Mehrjou, Si Kai Lee, Anant Raj
- abstract@[open-review](#): We present a novel algorithm for non-linear instrumental variable (IV) regression, DualIV, which simplifies traditional two-stage methods via a dual formulation. Inspired by problems in stochastic programming, we show that two-stage procedures for non-linear IV regression can be reformulated as a convex-concave saddle-point problem. Our formulation enables us to circumvent the first-stage regression which is a potential bottleneck in real-world applications. We develop a simple kernel-based algorithm with an analytic solution based on this formulation. Empirical results show that we are competitive to existing, more complicated algorithms for non-linear instrumental variable regression.

[Stochastic Gradient Descent in Correlated Settings: A Study on Gaussian Processes](#)

- Hao Chen, Lili Zheng, Raed AL Kontar, Garvesh Raskutti
- abstract@[open-review](#): Stochastic gradient descent (SGD) and its variants have established themselves as the go-to algorithms for large-scale machine learning problems with independent samples due to their generalization performance and intrinsic computational advantage. However, the fact that the stochastic gradient is a biased estimator of the full gradient with correlated samples has led to the lack of theoretical understanding of how SGD behaves under correlated settings and hindered its use in such cases. In this paper, we focus on the Gaussian process (GP) and take a step forward towards breaking the barrier by proving minibatch SGD converges to a critical point of the full loss function, and recovers model hyperparameters with rate $\mathcal{O}(\frac{1}{K})$ up to a statistical error term depending on the minibatch size. Numerical studies on both simulated and real datasets demonstrate that minibatch SGD has better generalization over state-of-the-art GP methods while reducing the computational burden and opening a new, previously unexplored, data size regime for GPs.

[Interventional Few-Shot Learning](#)

- Zhongqi Yue, Hanwang Zhang, Qianru Sun, Xian-Sheng Hua
- abstract@[open-review](#): We uncover an ever-overlooked deficiency in the prevailing Few-Shot Learning (FSL) methods: the pre-trained knowledge is indeed a confounder that limits the performance. This finding is rooted from our causal assumption: a Structural Causal Model (SCM) for the causalities among the pre-trained knowledge, sample features, and labels. Thanks to it, we propose a novel FSL paradigm: Interventional Few-Shot Learning (IFSL). Specifically, we develop three effective IFSL algorithmic implementations based on the backdoor adjustment, which is essentially a causal intervention towards the SCM of many-shot learning: the upper-bound of FSL in a causal view. It is worth noting that the contribution of IFSL is orthogonal to existing fine-tuning and meta-learning based FSL methods, hence IFSL can improve all of them, achieving a new 1/5-shot state-of-the-art on miniImageNet, tieredImageNet, and cross-domain CUB. Code is released at <https://github.com/yue-zhongqi/ifsl>.

[Minimax Value Interval for Off-Policy Evaluation and Policy Optimization](#)

- Nan Jiang, Jiawei Huang
- abstract@[open-review](#): We study minimax methods for off-policy evaluation (OPE) using value functions and marginalized importance weights. Despite that they hold promises of overcoming the exponential variance in traditional importance sampling, several key problems remain: (1) They require function approximation and are generally biased. For the sake of trustworthy OPE, is there anyway to quantify the biases? (2) They are split into two styles (``weight-learning'' vs ``value-learning''). Can we unify them? In this paper we answer both questions positively. By slightly altering the derivation of previous methods (one from each style), we unify them into a single value interval that comes with a special type of double robustness: when either the value-function or the importance-weight class is well specified, the interval is valid and its length quantifies the misspecification of the other class. Our interval also provides a unified view of and new insights to some recent methods, and we further explore the implications of our results on exploration and exploitation in off-policy policy optimization with insufficient data coverage.

[Biased Stochastic First-Order Methods for Conditional Stochastic Optimization and Applications in Meta Learning](#)

- Yifan Hu, Siqi Zhang, Xin Chen, Niao He
- abstract@[open-review](#): Conditional stochastic optimization covers a variety of applications ranging from invariant learning and causal inference to meta-learning. However, constructing unbiased gradient estimators for such problems is challenging due to the composition structure. As an alternative, we propose a biased stochastic gradient descent (BSGD) algorithm and study the bias-variance tradeoff under different structural assumptions. We establish the sample complexities of BSGD for strongly convex, convex, and weakly convex objectives under smooth and non-smooth conditions. Our lower bound analysis shows that the sample complexities of BSGD cannot be improved for general convex objectives and nonconvex objectives except for smooth nonconvex objectives with Lipschitz continuous gradient estimator. For this special setting, we propose an accelerated algorithm called biased SpiderBoost (BSpiderBoost) that matches the lower bound complexity. We further conduct numerical experiments on invariant logistic regression and model-agnostic meta-learning to illustrate the performance of BSGD and BSpiderBoost.

[ShiftAddNet: A Hardware-Inspired Deep Network](#)

- Haoran You, Xiaohan Chen, Yongan Zhang, Chaojian Li, Sicheng Li, Zihao Liu, Zhangyang Wang, Yingyan Lin

- abstract@[open-review](#): Multiplication (e.g., convolution) is arguably a cornerstone of modern deep neural networks (DNNs). However, intensive multiplications cause expensive resource costs that challenge DNNs' deployment on resource-constrained edge devices, driving several attempts for multiplication-less deep networks. This paper presented ShiftAddNet, whose main inspiration is drawn from a common practice in energy-efficient hardware implementation, that is, multiplication can be instead performed with additions and logical bit-shifts. We leverage this idea to explicitly parameterize deep networks in this way, yielding a new type of deep network that involves only bit-shift and additive weight layers. This hardware-inspired ShiftAddNet immediately leads to both energy-efficient inference and training, without compromising the expressive capacity compared to standard DNNs. The two complementary operation types (bit-shift and add) additionally enable finer-grained control of the model's learning capacity, leading to more flexible trade-off between accuracy and (training) efficiency, as well as improved robustness to quantization and pruning. We conduct extensive experiments and ablation studies, all backed up by our FPGA-based ShiftAddNet implementation and energy measurements. Compared to existing DNNs or other multiplication-less models, ShiftAddNet aggressively reduces over 80% hardware-quantified energy cost of DNNs training and inference, while offering comparable or better accuracies. Codes and pre-trained models are available at <https://github.com/RICE-EIC/ShiftAddNet>.

[Network-to-Network Translation with Conditional Invertible Neural Networks](#)

- Robin Rombach, Patrick Esser, Bjorn Ommer
- abstract@[open-review](#): Given the ever-increasing computational costs of modern machine learning models, we need to find new ways to reuse such expert models and thus tap into the resources that have been invested in their creation. Recent work suggests that the power of these massive models is captured by the representations they learn. Therefore, we seek a model that can relate between different existing representations and propose to solve this task with a conditionally invertible network. This network demonstrates its capability by (i) providing generic transfer between diverse domains, (ii) enabling controlled content synthesis by allowing modification in other domains, and (iii) facilitating diagnosis of existing representations by translating them into interpretable domains such as images. Our domain transfer network can translate between fixed representations without having to learn or finetune them. This allows users to utilize various existing domain-specific expert models from the literature that had been trained with extensive computational resources. Experiments on diverse conditional image synthesis tasks, competitive image modification results and experiments on image-to-image and text-to-image generation demonstrate the generic applicability of our approach. For example, we translate between BERT and BigGAN, state-of-the-art text and image models to provide text-to-image generation, which neither of both experts can perform on their own.

[Intra-Processing Methods for Debiasing Neural Networks](#)

- Yash Savani, Colin White, Naveen Sundar Govindarajulu
- abstract@[open-review](#): As deep learning models become tasked with more and more decisions that impact human lives, such as criminal recidivism, loan repayment, and face recognition for law enforcement, bias is becoming a growing concern. Debiasing algorithms are typically split into three paradigms: pre-processing, in-processing, and post-processing. However, in computer vision or natural language applications, it is common to start with a large generic model and then fine-tune to a specific use-case. Pre- or in-processing methods would require retraining the entire model from scratch, while post-processing methods only have black-box access to the model, so they do not leverage the weights of the trained model. Creating debiasing algorithms specifically for this fine-tuning use-case has largely been neglected. In this work, we initiate the study of a new paradigm in debiasing research, intra-processing, which sits between in-processing and post-processing methods. Intra-processing methods are designed specifically to debias large models which have been trained on a generic dataset, and fine-tuned on a more specific task. We show how to repurpose existing in-processing methods for this use-case, and we also propose three baseline algorithms: random perturbation, layerwise optimization, and adversarial debiasing. We evaluate these methods across three popular datasets from the AIF360 toolkit, as well as on the CelebA faces dataset. Our code is available at https://github.com/abacusai/intraprocessing_debiasing.

[Finding Second-Order Stationary Points Efficiently in Smooth Nonconvex Linearly Constrained Optimization Problems](#)

- Songtao Lu, Meisam Razaviyayn, Bo Yang, Kejun Huang, Mingyi Hong
- abstract@[open-review](#): This paper proposes two efficient algorithms for computing approximate second-order stationary points (SOSPs) of problems with generic smooth non-convex objective functions and generic linear constraints. While finding (approximate) SOSPs for the class of smooth non-convex linearly constrained problems is computationally intractable, we show that generic problem instances in this class can be solved efficiently. Specifically, for a generic problem instance, we show that certain strict complementarity (SC) condition holds for all Karush-Kuhn-Tucker (KKT) solutions. Based on this condition, we design an algorithm named Successive Negative-curvature grAdient Projection (SNAP), which performs either conventional gradient projection or some negative curvature-based projection steps to find SOSPs. SNAP is a second-order algorithm that requires $\widetilde{O}(\max\{1/\epsilon^2 G, 1/\epsilon^3 H\})$ iterations to compute an (ϵ_G, ϵ_H) -SOSP, where \widetilde{O} hides the iteration complexity for eigenvalue-decomposition. Building on SNAP, we propose a first-order algorithm, named SNAP^{A+}, that requires $\mathcal{O}(1/\epsilon^{2.5})$ iterations to compute $(\epsilon, \sqrt{\epsilon})$ -SOSP. The per-iteration computational complexities of our algorithms are polynomial in the number of constraints and problem dimension. To the best of our knowledge, this is the first time that first-order algorithms with polynomial per-iteration complexity and global sublinear rate are designed to find SOSPs of the important class of non-convex problems with linear constraints (almost surely).

[Model-based Policy Optimization with Unsupervised Model Adaptation](#)

- Jian Shen, Han Zhao, Weinan Zhang, Yong Yu
- abstract@[open-review](#): Model-based reinforcement learning methods learn a dynamics model with real data sampled from the environment and leverage it to generate simulated data to derive an agent. However, due to the potential distribution mismatch between simulated data and real data, this could lead to degraded performance. Despite much effort being devoted to reducing this distribution mismatch, existing methods fail to solve it explicitly. In this paper, we investigate how to bridge the gap between real and simulated data due to inaccurate model estimation for better policy optimization. To begin with, we first derive a lower bound of the expected return, which naturally inspires a bound maximization algorithm by aligning the simulated and real data distributions. To this end, we propose a novel model-based reinforcement learning framework AMPO, which introduces unsupervised model adaptation to minimize the integral probability metric (IPM) between feature distributions from real and simulated data. Instantiating our framework with Wasserstein-1 distance gives a practical model-based approach. Empirically, our approach achieves state-of-the-art performance in terms of sample efficiency on a range of continuous control benchmark tasks.

[Implicit Regularization and Convergence for Weight Normalization](#)

- Xiaoxia Wu, Edgar Dobriban, Tongzheng Ren, Shanshan Wu, Zhiyuan Li, Suriya Gunasekar, Rachel Ward, Qiang Liu
- abstract@[open-review](#): Normalization methods such as batch, weight, instance, and layer normalization are commonly used in modern machine learning. Here, we study the weight normalization (WN) method \cite{salimans2016weight} and a variant called reparametrized projected gradient descent (rPGD) for overparametrized least squares regression and some more general loss functions. WN and rPGD reparametrize the weights with a scale $\|g\|$ and a unit vector such that the objective function becomes $\|\mathbf{w}\|_{\text{non-convex}}$. We show that this non-convex formulation has beneficial regularization effects compared to gradient descent on the original objective. These methods adaptively regularize the weights and converge linearly close to the minimum ℓ_2 norm solution even for initializations far from zero. For certain two-phase variants, they can converge to the min norm solution. This is different from the behavior of gradient descent, which only converges to the min norm solution when started at zero, and thus more sensitive to initialization.

[Geometric All-way Boolean Tensor Decomposition](#)

- Changlin Wan, Wennan Chang, Tong Zhao, Sha Cao, Chi Zhang
- abstract@[open-review](#): Boolean tensor has been broadly utilized in representing high dimensional logical data collected on spatial, temporal and/or other relational domains. Boolean Tensor Decomposition (BTD) factorizes a binary tensor into the Boolean sum of multiple rank-1 tensors, which is an NP-hard problem. Existing BTD methods have been limited by their high computational cost, in applications to large scale or higher order tensors. In this work, we presented a computationally efficient BTD algorithm, namely Geometric Expansion for all-order Tensor Factorization (GETF), that sequentially identifies the rank-1 basis components for a tensor from a geometric perspective. We conducted rigorous theoretical analysis on the validity as well as algorithmic efficiency of GETF in decomposing all-order tensor. Experiments on both synthetic and real-world data demonstrated that GETF has significantly improved performance in reconstruction accuracy, extraction of latent structures and it is an order of magnitude faster than other state-of-the-art methods.

[Modular Meta-Learning with Shrinkage](#)

- Yutian Chen, Abram L. Friesen, Feryal Behbahani, Arnaud Doucet, David Budden, Matthew Hoffman, Nando de Freitas
- abstract@[open-review](#): Many real-world problems, including multi-speaker text-to-speech synthesis, can greatly benefit from the ability to meta-learn large models with only a few task-specific components. Updating only these task-specific modules then allows the model to be adapted to low-data tasks for as many steps as necessary without risking overfitting. Unfortunately, existing meta-learning methods either do not scale to long adaptation or else rely on handcrafted task-specific architectures. Here, we propose a meta-learning approach that obviates the need for this often sub-optimal hand-selection. In particular, we develop general techniques based on Bayesian shrinkage to automatically discover and learn both task-specific and general reusable modules. Empirically, we demonstrate that our method discovers a small set of meaningful task-specific modules and outperforms existing meta-learning approaches in domains like few-shot text-to-speech that have little task data and long adaptation horizons. We also show that existing meta-learning methods including MAML, iMAML, and Reptile emerge as special cases of our method.

[A/B Testing in Dense Large-Scale Networks: Design and Inference](#)

- Preetam Nandy, Kinjal Basu, Shaunak Chatterjee, Ye Tu
- abstract@[open-review](#): Design of experiments and estimation of treatment effects in large-scale networks, in the presence of strong interference, is a challenging and important problem. Most existing methods' performance deteriorates as the density of the network increases. In this paper, we present a novel strategy for accurately estimating the causal effects of a class of treatments in a dense large-scale network. First, we design an approximate randomized controlled experiment by solving an optimization problem to allocate treatments in the presence of competition among neighboring nodes. Then we apply an importance sampling adjustment to correct for any leftover bias (from the approximation) in estimating average treatment effects. We provide theoretical guarantees, verify robustness in a simulation study, and validate the scalability and usefulness of our procedure in a real-world experiment on a large social network.

[What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation](#)

- Vitaly Feldman, Chiyuan Zhang
- abstract@[open-review](#): Deep learning algorithms are well-known to have a propensity for fitting the training data very well and often fit even outliers and mislabeled data points. Such fitting requires memorization of training data labels, a phenomenon that has attracted significant research interest but has not been given a compelling explanation so far. A recent work of Feldman (2019) proposes a theoretical explanation for this phenomenon based on a combination of two insights. First, natural image and data distributions are (informally) known to be long-tailed, that is have a significant fraction of rare and atypical examples. Second, in a simple theoretical model such memorization is necessary for achieving close-to-optimal generalization error when the data distribution is long-tailed. However, no direct empirical evidence for this explanation or even an approach for obtaining such evidence were given. In this work we design experiments to test the key ideas in this theory. The experiments require estimation of the influence of each training example on the accuracy at each test example as well as memorization values of training examples. Estimating these quantities directly is computationally prohibitive but we show that closely-related subsampled influence and memorization values can be estimated much more efficiently. Our experiments demonstrate the significant benefits of memorization for generalization on several standard benchmarks. They also provide quantitative and visually compelling evidence for the theory put forth in Feldman (2019).

[Partially View-aligned Clustering](#)

- Zhenyu Huang, Peng Hu, Joey Tianyi Zhou, Jiancheng Lv, Xi Peng
- abstract@[open-review](#): In this paper, we study one challenging issue in multi-view data clustering. To be specific, for two data matrices $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ corresponding to two views, we do not assume that $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are fully aligned in row-wise. Instead, we assume that only a small portion of the matrices has established the correspondence in advance. Such a partially view-aligned problem (PVP) could lead to the intensive labor of capturing or establishing the aligned multi-view data, which has less been touched so far to the best of our knowledge. To solve this practical and challenging problem, we propose a novel multi-view clustering method termed partially view-aligned clustering (PVC). To be specific, PVC proposes to use a differentiable surrogate of the non-differentiable Hungarian algorithm and recasts it as a pluggable module. As a result, the category-level correspondence of the unaligned data could be established in a latent space learned by a neural network, while learning a common space across different views using the ``aligned'' data. Extensive experimental results show promising results of our method in clustering partially view-aligned data.

[Partial Optimal Transport with applications on Positive-Unlabeled Learning](#)

- Laetitia Chapel, Mokhtar Z. Alaya, Gilles Gasso
- abstract@[open-review](#): Classical optimal transport problem seeks a transportation map that preserves the total mass between two probability distributions, requiring their masses to be equal. This may be too restrictive in some applications such as color or shape matching, since the distributions may have arbitrary masses and/or only a fraction of the total mass has to be transported. In this paper, we address the partial Wasserstein and Gromov-Wasserstein problems and propose exact algorithms to solve them. We showcase the new formulation in a positive-unlabeled (PU) learning application. To the best of our knowledge, this is the first application of optimal transport in this context and we first highlight that partial Wasserstein-based metrics prove effective in usual PU learning settings. We then demonstrate that partial Gromov-Wasserstein metrics are efficient in scenarios in which the samples from the positive and the unlabeled datasets come from different domains or have different features.

[Toward the Fundamental Limits of Imitation Learning](#)

- Nived Rajaraman, Lin Yang, Jiantao Jiao, Kannan Ramchandran
- abstract@[open-review](#): Imitation learning (IL) aims to mimic the behavior of an expert policy in a sequential decision-making problem given only demonstrations. In this paper, we focus on understanding the minimax statistical limits of IL in episodic Markov Decision Processes (MDPs). We first consider the setting where the learner is provided a dataset of S expert trajectories ahead of time, and cannot interact with the MDP. Here, we show that the policy which mimics the expert whenever possible is in expectation $\frac{1}{H^2 \log(N)}N$ suboptimal compared to the value of the expert, even when the expert plays a stochastic policy. Here S is the state space and H is the length of the episode.

Furthermore, we establish a suboptimality lower bound of $\|\mathcal{G}^t\| \mathcal{H}^2 / N$ which applies even if the expert is constrained to be deterministic, or if the learner is allowed to actively query the expert at visited states while interacting with the MDP for N episodes. To our knowledge, this is the first algorithm with suboptimality having no dependence on the number of actions, under no additional assumptions. We then propose a novel algorithm based on minimum-distance functionals in the setting where the transition model is given and the expert is deterministic. The algorithm is suboptimal by $\|\mathcal{S}\| H^{3/2} / N$, matching our lower bound up to a \sqrt{H} factor, and breaks the $\mathcal{O}(H^2)$ error compounding barrier of IL.

[Logarithmic Pruning is All You Need](#)

- Laurent Orseau, Marcus Hutter, Omar Rivasplata
- abstract@[open-review](#): The Lottery Ticket Hypothesis is a conjecture that every large neural network contains a subnetwork that, when trained in isolation, achieves comparable performance to the large network. An even stronger conjecture has been proven recently: Every sufficiently overparameterized network contains a subnetwork that, even without training, achieves comparable accuracy to the trained large network. This theorem, however, relies on a number of strong assumptions and guarantees a polynomial factor on the size of the large network compared to the target function. In this work, we remove the most limiting assumptions of this previous work while providing significantly tighter bounds: the overparameterized network only needs a logarithmic factor (in all variables but depth) number of neurons per weight of the target subnetwork.

[Hold me tight! Influence of discriminative features on deep network boundaries](#)

- Guillermo Ortiz-Jimenez, Apostolos Modas, Seyed-Mohsen Moosavi, Pascal Frossard
- abstract@[open-review](#): Important insights towards the explainability of neural networks reside in the characteristics of their decision boundaries. In this work, we borrow tools from the field of adversarial robustness, and propose a new perspective that relates dataset features to the distance of samples to the decision boundary. This enables us to carefully tweak the position of the training samples and measure the induced changes on the boundaries of CNNs trained on large-scale vision datasets. We use this framework to reveal some intriguing properties of CNNs. Specifically, we rigorously confirm that neural networks exhibit a high invariance to non-discriminative features, and show that the decision boundaries of a DNN can only exist as long as the classifier is trained with some features that hold them together. Finally, we show that the construction of the decision boundary is extremely sensitive to small perturbations of the training samples, and that changes in certain directions can lead to sudden invariances in the orthogonal ones. This is precisely the mechanism that adversarial training uses to achieve robustness.

[Learning from Mixtures of Private and Public Populations](#)

- Raef Bassily, Shay Moran, Anupama Nandi
- abstract@[open-review](#): We initiate the study of a new model of supervised learning under privacy constraints. Imagine a medical study where a dataset is sampled from a population of both healthy and unhealthy individuals. Suppose healthy individuals have no privacy concerns (in such case, we call their data ``public'') while the unhealthy individuals desire stringent privacy protection for their data. In this example, the population (data distribution) is a mixture of private (unhealthy) and public (healthy) sub-populations that could be very different.

Inspired by the above example, we consider a model in which the population \mathcal{C} is a mixture of two possibly distinct sub-populations: a private sub-population D_{priv} of private and sensitive data, and a public sub-population D_{pub} of data with no privacy concerns. Each example drawn from \mathcal{C} is assumed to contain a privacy-status bit that indicates whether the example is private or public. The goal is to design a learning algorithm that satisfies differential privacy only with respect to the private examples.

Prior works in this context assumed a homogeneous population where private and public data arise from the same distribution, and in particular designed solutions which exploit this assumption. We demonstrate how to circumvent this assumption by considering, as a case study, the problem of learning linear classifiers in R^d . We show that in the case where the privacy status is correlated with the target label (as in the above example), linear classifiers in R^d can be learned, in the agnostic as well as the realizable setting, with sample complexity which is comparable to that of the classical (non-private) PAC-learning. It is known that this task is impossible if all the data is considered private.

[Adversarial Weight Perturbation Helps Robust Generalization](#)

- Dongxian Wu, Shu-Tao Xia, Yisen Wang
- abstract@[open-review](#): The study on improving the robustness of deep neural networks against adversarial examples grows rapidly in recent years. Among them, adversarial training is the most promising one, which flattens the loss landscape (loss change with respect to input) via training on adversarially perturbed examples. However, how the widely used weight loss landscape (loss change with respect to weight) performs in adversarial training is rarely explored. In this paper, we investigate the weight loss landscape from a new perspective, and identify a clear correlation between the flatness of weight loss landscape and robust generalization gap. Several well-recognized adversarial training improvements, such as early stopping, designing new objective functions, or leveraging unlabeled data, all implicitly flatten the weight loss landscape. Based on these observations, we propose a simple yet effective Adversarial Weight Perturbation (AWP) to explicitly regularize the flatness of weight loss landscape, forming a double-perturbation mechanism in the adversarial training framework that adversarially perturbs both inputs and weights. Extensive experiments demonstrate that AWP indeed brings flatter weight loss landscape and can be easily incorporated into various existing adversarial training methods to further boost their adversarial robustness.

[Stateful Posted Pricing with Vanishing Regret via Dynamic Deterministic Markov Decision Processes](#)

- Yuval Emek, Ron Lavi, Rad Niazadeh, Yangguang Shi
- abstract@[open-review](#): In this paper, a rather general online problem called dynamic resource allocation with capacity constraints (DRACC) is introduced and studied in the realm of posted price mechanisms. This problem subsumes several applications of stateful pricing, including but not limited to posted prices for online job scheduling and matching over a dynamic bipartite graph. As the existing online learning techniques do not yield vanishing-regret mechanisms for this problem, we develop a novel online learning framework defined over deterministic Markov decision processes with dynamic state transition and reward functions. We then prove that if the Markov decision process is guaranteed to admit an oracle that can simulate any given policy from any initial state with bounded loss --- a condition that is satisfied in the DRACC problem --- then the online learning problem can be solved with vanishing regret. Our proof technique is based on a reduction to online learning with switching cost, in which an online decision maker incurs an extra cost every time she switches from one arm to another. We formally demonstrate this connection and further show how DRACC can be used in our proposed applications of stateful pricing.

[Adversarial Self-Supervised Contrastive Learning](#)

- Minseon Kim, Jihoon Tack, Sung Ju Hwang
- abstract@[open-review](#): Existing adversarial learning approaches mostly use class labels to generate adversarial samples that lead to incorrect predictions, which are then used to augment the training of the model for improved robustness. While some recent works propose semi-supervised adversarial learning methods that utilize unlabeled data, they still require class labels. However, do we really need class labels at all, for adversarially robust training of deep neural networks? In this paper, we propose a novel adversarial attack for unlabeled data, which makes the model confuse the instance-level identities of

the perturbed data samples. Further, we present a self-supervised contrastive learning framework to adversarially train a robust neural network without labeled data, which aims to maximize the similarity between a random augmentation of a data sample and its instance-wise adversarial perturbation. We validate our method, Robust Contrastive Learning (RoCL), on multiple benchmark datasets, on which it obtains comparable robust accuracy over state-of-the-art supervised adversarial learning methods, and significantly improved robustness against the \emph{black box} and unseen types of attacks. Moreover, with further joint fine-tuning with supervised adversarial loss, RoCL obtains even higher robust accuracy over using self-supervised learning alone. Notably, RoCL also demonstrate impressive results in robust transfer learning.

[Normalizing Kalman Filters for Multivariate Time Series Analysis](#)

- Emmanuel de BÃ©zenac, Syama Sundar Rangapuram, Konstantinos Benidis, Michael Bohlke-Schneider, Richard Kurle, Lorenzo Stella, Hilaf Hasson, Patrick Gallinari, Tim Januschowski
- abstract@[open-review](#): This paper tackles the modelling of large, complex and multivariate time series panels in a probabilistic setting. To this extent, we present a novel approach reconciling classical state space models with deep learning methods. By augmenting state space models with normalizing flows, we mitigate imprecisions stemming from idealized assumptions in state space models. The resulting model is highly flexible while still retaining many of the attractive properties of state space models, e.g., uncertainty and observation errors are properly accounted for, inference is tractable, sampling is efficient, good generalization performance is observed, even in low data regimes. We demonstrate competitiveness against state-of-the-art deep learning methods on the tasks of forecasting real world data and handling varying levels of missing data.

[Learning to summarize with human feedback](#)

- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, Paul F. Christiano
- abstract@[open-review](#): As language models become more powerful, training and evaluation are increasingly bottlenecked by the data and metrics used for a particular task. For example, summarization models are often trained to predict human reference summaries and evaluated using ROUGE, but both of these metrics are rough proxies for what we really care about---summary quality. In this work, we show that it is possible to significantly improve summary quality by training a model to optimize for human preferences. We collect a large, high-quality dataset of human comparisons between summaries, train a model to predict the human-preferred summary, and use that model as a reward function to fine-tune a summarization policy using reinforcement learning. We apply our method to a version of the TL;DR dataset of Reddit posts and find that our models significantly outperform both human reference summaries and much larger models fine-tuned with supervised learning alone. Our models also transfer to CNN/DM news articles, producing summaries nearly as good as the human reference without any news-specific fine-tuning. We conduct extensive analyses to understand our human feedback dataset and fine-tuned models. We establish that our reward model generalizes to new datasets, and that optimizing our reward model results in better summaries than optimizing ROUGE according to humans. We hope the evidence from our paper motivates machine learning researchers to pay closer attention to how their training loss affects the model behavior they actually want.

[Fourier Spectrum Discrepancies in Deep Network Generated Images](#)

- Tarik Dzanic, Karan Shah, Freddie Witherden
- abstract@[open-review](#): Advancements in deep generative models such as generative adversarial networks and variational autoencoders have resulted in the ability to generate realistic images that are visually indistinguishable from real images which raises concerns about their potential malicious usage. In this paper, we present an analysis of the high-frequency Fourier modes of real and deep network generated images and show that deep network generated images share an observable, systematic shortcoming in replicating the attributes of these high-frequency modes. Using this, we propose a novel detection method based on the frequency spectrum of the images which is able to achieve an accuracy of up to 99.2% in classifying real and deep network generated images from various GAN and VAE architectures on a dataset of 5000 images with as few as 8 training examples. Furthermore, we show the impact of image transformations such as compression, cropping, and resolution reduction on the classification accuracy and suggest a method for modifying the high-frequency attributes of deep network generated images to mimic real images.

[Lamina-specific neuronal properties promote robust, stable signal propagation in feedforward networks](#)

- Dongqi Han, Erik De Schutter, Sungho Hong
- abstract@[open-review](#): Feedforward networks (FFN) are ubiquitous structures in neural systems and have been studied to understand mechanisms of reliable signal and information transmission. In many FFNs, neurons in one layer have intrinsic properties that are distinct from those in their pre-/postsynaptic layers, but how this affects network-level information processing remains unexplored. Here we show that layer-to-layer heterogeneity arising from lamina-specific cellular properties facilitates signal and information transmission in FFNs. Specifically, we found that signal transformations, made by each layer of neurons on an input-driven spike signal, demodulate signal distortions introduced by preceding layers. This mechanism boosts information transfer carried by a propagating spike signal, and thereby supports reliable spike signal and information transmission in a deep FFN. Our study suggests that distinct cell types in neural circuits, performing different computational functions, facilitate information processing on the whole.

[Learning Dynamic Belief Graphs to Generalize on Text-Based Games](#)

- Ashutosh Adhikari, Xingdi Yuan, Marc-Alexandre CÃ´tÃ©, MikulÃ¡Å¡ Zelinka, Marc-Antoine Rondeau, Romain Laroche, Pascal Poupart, Jian Tang, Adam Trischler, Will Hamilton
- abstract@[open-review](#): Playing text-based games requires skills in processing natural language and sequential decision making. Achieving human-level performance on text-based games remains an open challenge, and prior research has largely relied on hand-crafted structured representations and heuristics. In this work, we investigate how an agent can plan and generalize in text-based games using graph-structured representations learned end-to-end from raw text. We propose a novel graph-aided transformer agent (GATA) that infers and updates latent belief graphs during planning to enable effective action selection by capturing the underlying game dynamics. GATA is trained using a combination of reinforcement and self-supervised learning. Our work demonstrates that the learned graph-based representations help agents converge to better policies than their text-only counterparts and facilitate effective generalization across game configurations. Experiments on 500+ unique games from the TextWorld suite show that our best agent outperforms text-based baselines by an average of 24.2%.

[Triple descent and the two kinds of overfitting: where & why do they appear?](#)

- StÃ©phane d'Ascoli, Levent Sagun, Giulio Biroli
- abstract@[open-review](#): A recent line of research has highlighted the existence of a ``double descent'' phenomenon in deep learning, whereby increasing the number of training examples N causes the generalization error of neural networks to peak when N is of the same order as the number of parameters P . In earlier works, a similar phenomenon was shown to exist in simpler models such as linear regression, where the peak instead occurs when N is equal to the input dimension D . Since both peaks coincide with the interpolation threshold, they are often conflated in the literature. In this paper, we show that despite their apparent similarity, these two scenarios are inherently different. In fact, both peaks can co-exist when neural networks are applied to noisy regression tasks. The relative size of the peaks is then governed by the degree of nonlinearity of the activation function. Building on recent developments in the analysis of random feature models, we provide a theoretical ground for this sample-wise triple descent. As shown previously, the nonlinear peak at $N=P$ is a true divergence caused by the extreme sensitivity of the output function to both the noise corrupting the labels and the initialization of the random features (or the weights in neural networks). This peak survives in the absence of noise, but can be suppressed by regularization. In contrast, the linear peak at $N=D$ is solely due to overfitting the noise in the labels, and forms earlier during training. We show that this peak is implicitly regularized by

the nonlinearity, which is why it only becomes salient at high noise and is weakly affected by explicit regularization. Throughout the paper, we compare the analytical results obtained in the random feature model with the outcomes of numerical experiments involving realistic neural networks.

[Multimodal Graph Networks for Compositional Generalization in Visual Question Answering](#)

- Raeid Saqr, Karthik Narasimhan
- abstract@[open-review](#): Compositional generalization is a key challenge in grounding natural language to visual perception. While deep learning models have achieved great success in multimodal tasks like visual question answering, recent studies have shown that they fail to generalize to new inputs that are simply an unseen combination of those seen in the training distribution. In this paper, we propose to tackle this challenge by employing neural factor graphs to induce a tighter coupling between concepts in different modalities (e.g. images and text). Graph representations are inherently compositional in nature and allow us to capture entities, attributes and relations in a scalable manner. Our model first creates a multimodal graph, processes it with a graph neural network to induce a factor correspondence matrix, and then outputs a symbolic program to predict answers to questions. Empirically, our model achieves close to perfect scores on a caption truth prediction problem and state-of-the-art results on the recently introduced CLOSURE dataset, improving on the mean overall accuracy across seven compositional templates by 4.77% over previous approaches.

[Learning Graph Structure With A Finite-State Automaton Layer](#)

- Daniel Johnson, Hugo Larochelle, Daniel Tarlow
- abstract@[open-review](#): Graph-based neural network models are producing strong results in a number of domains, in part because graphs provide flexibility to encode domain knowledge in the form of relational structure (edges) between nodes in the graph. In practice, edges are used both to represent intrinsic structure (e.g., abstract syntax trees of programs) and more abstract relations that aid reasoning for a downstream task (e.g., results of relevant program analyses). In this work, we study the problem of learning to derive abstract relations from the intrinsic graph structure. Motivated by their power in program analyses, we consider relations defined by paths on the base graph accepted by a finite-state automaton. We show how to learn these relations end-to-end by relaxing the problem into learning finite-state automata policies on a graph-based POMDP and then training these policies using implicit differentiation. The result is a differentiable Graph Finite-State Automaton (GFSA) layer that adds a new edge type (expressed as a weighted adjacency matrix) to a base graph. We demonstrate that this layer can find shortcuts in grid-world graphs and reproduce simple static analyses on Python programs. Additionally, we combine the GFSA layer with a larger graph-based model trained end-to-end on the variable misuse program understanding task, and find that using the GFSA layer leads to better performance than using hand-engineered semantic edges or other baseline methods for adding learned edge types.

[A Universal Approximation Theorem of Deep Neural Networks for Expressing Probability Distributions](#)

- Yulong Lu, Jianfeng Lu
- abstract@[open-review](#): This paper studies the universal approximation property of deep neural networks for representing probability distributions. Given a target distribution π and a source distribution p_z both defined on \mathbb{R}^d , we prove under some assumptions that there exists a deep neural network $g: \mathbb{R}^d \rightarrow \mathbb{R}^d$ with ReLU activation such that the push-forward measure $(\nabla g)_* p_z$ of p_z under the map ∇g is arbitrarily close to the target measure π . The closeness are measured by three classes of integral probability metrics between probability distributions: 1 -Wasserstein distance, maximum mean distance (MMD) and kernelized Stein discrepancy (KSD). We prove upper bounds for the size (width and depth) of the deep neural network in terms of the dimension d and the approximation error ε with respect to the three discrepancies. In particular, the size of neural network can grow exponentially in d when 1 -Wasserstein distance is used as the discrepancy, whereas for both MMD and KSD the size of neural network only depends on d at most polynomially. Our proof relies on convergence estimates of empirical measures under aforementioned discrepancies and semi-discrete optimal transport.

[Unsupervised object-centric video generation and decomposition in 3D](#)

- Paul Henderson, Christoph H. Lampert
- abstract@[open-review](#): A natural approach to generative modeling of videos is to represent them as a composition of moving objects. Recent works model a set of 2D sprites over a slowly-varying background, but without considering the underlying 3D scene that gives rise to them. We instead propose to model a video as the view seen while moving through a scene with multiple 3D objects and a 3D background. Our model is trained from monocular videos without any supervision, yet learns to generate coherent 3D scenes containing several moving objects. We conduct detailed experiments on two datasets, going beyond the visual complexity supported by state-of-the-art generative approaches. We evaluate our method on depth-prediction and 3D object detection---tasks which cannot be addressed by those earlier works---and show it out-performs them even on 2D instance segmentation and tracking.

[Domain Generalization for Medical Imaging Classification with Linear-Dependency Regularization](#)

- Haoliang Li, Yufei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, Alex Kot
- abstract@[open-review](#): Recently, we have witnessed great progress in the field of medical imaging classification by adopting deep neural networks. However, the recent advanced models still require accessing sufficiently large and representative datasets for training, which is often unfeasible in clinically realistic environments. When trained on limited datasets, the deep neural network is lack of generalization capability, as the trained deep neural network on data within a certain distribution (e.g. the data captured by a certain device vendor or patient population) may not be able to generalize to the data with another distribution. In this paper, we introduce a simple but effective approach to improve the generalization capability of deep neural networks in the field of medical imaging classification. Motivated by the observation that the domain variability of the medical images is to some extent compact, we propose to learn a representative feature space through variational encoding with a novel linear-dependency regularization term to capture the shareable information among medical data collected from different domains. As a result, the trained neural network is expected to equip with better generalization capability to the ``unseen'' medical data. Experimental results on two challenging medical imaging classification tasks indicate that our method can achieve better cross-domain generalization capability compared with state-of-the-art baselines.

[Multi-label classification: do Hamming loss and subset accuracy really conflict with each other?](#)

- Guoqiang Wu, Jun Zhu
- abstract@[open-review](#): Various evaluation measures have been developed for multi-label classification, including Hamming Loss (HL), Subset Accuracy (SA) and Ranking Loss (RL). However, there is a gap between empirical results and the existing theories: 1) an algorithm often empirically performs well on some measure(s) while poorly on others, while a formal theoretical analysis is lacking; and 2) in small label space cases, the algorithms optimizing HL often have comparable or even better performance on the SA measure than those optimizing SA directly, while existing theoretical results show that SA and HL are conflicting measures. This paper provides an attempt to fill up this gap by analyzing the learning guarantees of the corresponding learning algorithms on both SA and HL measures. We show that when a learning algorithm optimizes HL with its surrogate loss, it enjoys an error bound for the HL measure independent of c (the number of labels), while the bound for the SA measure depends on at most $O(c)$. On the other hand, when directly optimizing SA with its surrogate loss, it has learning guarantees that depend on $O(\sqrt{c})$ for both HL and SA measures. This explains the observation that when the label space is not large, optimizing HL with its surrogate loss can have promising performance for SA. We further show that our techniques are applicable to analyze the learning guarantees of algorithms on other measures, such as RL. Finally, the theoretical analyses are supported by experimental results.

[A Novel Automated Curriculum Strategy to Solve Hard Sokoban Planning Instances](#)

- Dieqiao Feng, Carla P. Gomes, Bart Selman
- abstract@[open-review](#): In recent years, we have witnessed tremendous progress in deep reinforcement learning (RL) for tasks such as Go, Chess, video games, and robot control. Nevertheless, other combinatorial domains, such as AI planning, still pose considerable challenges for RL approaches. The key difficulty in those domains is that a positive reward signal becomes {\em exponentially rare} as the minimal solution length increases. So, an RL approach loses its training signal. There has been promising recent progress by using a curriculum-driven learning approach that is designed to solve a single hard instance. We present a novel {\em automated} curriculum approach that dynamically selects from a pool of unlabeled training instances of varying task complexity guided by our {\em difficulty quantum momentum} strategy. We show how the smoothness of the task hardness impacts the final learning results. In particular, as the size of the instance pool increases, the ``hardness gap'' decreases, which facilitates a smoother automated curriculum based learning process. Our automated curriculum approach dramatically improves upon the previous approaches. We show our results on Sokoban, which is a traditional PSPACE-complete planning problem and presents a great challenge even for specialized solvers. Our RL agent can solve hard instances that are far out of reach for any previous state-of-the-art Sokoban solver. In particular, our approach can uncover plans that require hundreds of steps, while the best previous search methods would take many years of computing time to solve such instances. In addition, we show that we can further boost the RL performance with an intricate coupling of our automated curriculum approach with a curiosity-driven search strategy and a graph neural net representation.

[Causal analysis of Covid-19 Spread in Germany](#)

- Atalanti Mastakouri, Bernhard SchÄlkopf
- abstract@[open-review](#): In this work, we study the causal relations among German regions in terms of the spread of Covid-19 since the beginning of the pandemic, taking into account the restriction policies that were applied by the different federal states. We loose a strictly formulated assumption for a causal feature selection method for time series data, robust to latent confounders, which we subsequently apply on Covid-19 case numbers. We present findings about the spread of the virus in Germany and the causal impact of restriction measures, discussing the role of various policies in containing the spread. Since our results are based on rather limited target time series (only the numbers of reported cases), care should be exercised in interpreting them. However, it is encouraging that already such limited data seems to contain causal signals. This suggests that as more data becomes available, our causal approach may contribute towards meaningful causal analysis of political interventions on the development of Covid-19, and thus also towards the development of rational and data-driven methodologies for choosing interventions.

[Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms](#)

- Thomas Berrett, Cristina Butucea
- abstract@[open-review](#): We find separation rates for testing multinomial or more general discrete distributions under the constraint of alpha-local differential privacy. We construct efficient randomized algorithms and test procedures, in both the case where only non-interactive privacy mechanisms are allowed and also in the case where all sequentially interactive privacy mechanisms are allowed. The separation rates are faster in the latter case. We prove general information theoretical bounds that allow us to establish the optimality of our algorithms among all pairs of privacy mechanisms and test procedures, in most usual cases. Considered examples include testing uniform, polynomially and exponentially decreasing distributions.

[Adaptive Gradient Quantization for Data-Parallel SGD](#)

- Fartash Faghri, Iman Tabrizian, Ilia Markov, Dan Alistarh, Daniel M. Roy, Ali Ramezani-Kebrya
- abstract@[open-review](#): Many communication-efficient variants of SGD use gradient quantization schemes. These schemes are often heuristic and fixed over the course of training. We empirically observe that the statistics of gradients of deep models change during the training. Motivated by this observation, we introduce two adaptive quantization schemes, ALQ and AMQ. In both schemes, processors update their compression schemes in parallel by efficiently computing sufficient statistics of a parametric distribution. We improve the validation accuracy by almost 2% on CIFAR-10 and 1% on ImageNet in challenging low-cost communication setups. Our adaptive methods are also significantly more robust to the choice of hyperparameters.

[Finite Continuum-Armed Bandits](#)

- Solenne Gaucher
- abstract@[open-review](#): We consider a situation where an agent has \$T\$ resources to be allocated to a larger number \$N\$ of actions. Each action can be completed at most once and results in a stochastic reward with unknown mean. The goal of the agent is to maximize her cumulative reward. Non trivial strategies are possible when side information on the actions is available, for example in the form of covariates. Focusing on a nonparametric setting, where the mean reward is an unknown function of a one-dimensional covariate, we propose an optimal strategy for this problem. Under natural assumptions on the reward function, we prove that the optimal regret scales as \$O(T^{1/3})\$ up to poly-logarithmic factors when the budget \$T\$ is proportional to the number of actions \$N\$. When \$T\$ becomes small compared to \$N\$, a smooth transition occurs. When the ratio \$T/N\$ decreases from a constant to \$N^{-1/3}\$, the regret increases progressively up to the \$O(T^{1/2})\$ rate encountered in continuum-armed bandits.

[Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies](#)

- Itai Gat, Idan Schwartz, Alexander Schwing, Tamir Hazan
- abstract@[open-review](#): Many recent datasets contain a variety of different data modalities, for instance, image, question, and answer data in visual question answering (VQA). When training deep net classifiers on those multi-modal datasets, the modalities get exploited at different scales, i.e., some modalities can more easily contribute to the classification results than others. This is suboptimal because the classifier is inherently biased towards a subset of the modalities. To alleviate this shortcoming, we propose a novel regularization term based on the functional entropy. Intuitively, this term encourages to balance the contribution of each modality to the classification result. However, regularization with the functional entropy is challenging. To address this, we develop a method based on the log-Sobolev inequality, which bounds the functional entropy with the functional-Fisher-information. Intuitively, this maximizes the amount of information that the modalities contribute. On the two challenging multi-modal datasets VQA-CPv2, and SocialIQ, we obtain state-of-the-art results while more uniformly exploiting the modalities. In addition, we demonstrate the efficacy of our method on Colored MNIST.

[Compact task representations as a normative model for higher-order brain activity](#)

- Severin Berger, Christian K. Machens
- abstract@[open-review](#): Higher-order brain areas such as the frontal cortices are considered essential for the flexible solution of tasks. However, the precise computational role of these areas is still debated. Indeed, even for the simplest of tasks, we cannot really explain how the measured brain activity, which evolves over time in complicated ways, relates to the task structure. Here, we follow a normative approach, based on integrating the principle of efficient coding with the framework of Markov decision processes (MDP). More specifically, we focus on MDPs whose state is based on action-observation histories, and we show how to compress the state space such that unnecessary redundancy is eliminated, while task-relevant information is preserved. We show that the efficiency of a state space representation depends on the (long-term) behavioural goal of the agent, and we distinguish between model-based and habitual agents. We apply our approach to simple tasks that require short-term memory, and we show that the efficient state space representations

reproduce the key dynamical features of recorded neural activity in frontal areas (such as ramping, sequentiality, persistence). If we additionally assume that neural systems are subject to accuracy-cost tradeoffs, we find a surprising match to neural data on a population level.

[Robust-Adaptive Control of Linear Systems: beyond Quadratic Costs](#)

- Edouard Leurent, Odalric-Ambrym Maillard, Denis Efimov
- abstract@[open-review](#): We consider the problem of robust and adaptive model predictive control (MPC) of a linear system, with unknown parameters that are learned along the way (adaptive), in a critical setting where failures must be prevented (robust). This problem has been studied from different perspectives by different communities. However, the existing theory deals only with the case of quadratic costs (the LQ problem), which limits applications to stabilisation and tracking tasks only. In order to handle more general (non-convex) costs that naturally arise in many practical problems, we carefully select and bring together several tools from different communities, namely non-asymptotic linear regression, recent results in interval prediction, and tree-based planning. Combining and adapting the theoretical guarantees at each layer is non trivial, and we provide the first end-to-end suboptimality analysis for this setting. Interestingly, our analysis naturally adapts to handle many models and combines with a data-driven robust model selection strategy, which enables to relax the modelling assumptions. Last, we strive to preserve tractability at any stage of the method, that we illustrate on two challenging simulated environments.

[Co-exposure Maximization in Online Social Networks](#)

- Sijing Tu, Cigdem Aslay, Aristides Gionis
- abstract@[open-review](#): Social media has created new ways for citizens to stay informed on societal matters and participate in political discourse. However, with its algorithmically-curated and virally-propagating content, social media has contributed further to the polarization of opinions by reinforcing users' existing viewpoints. An emerging line of research seeks to understand how content-recommendation algorithms can be re-designed to mitigate societal polarization amplified by social-media interactions. In this paper, we study the problem of allocating seed users to opposing campaigns: by drawing on the equal-time rule of political campaigning on traditional media, our goal is to allocate seed users to campaigners with the aim to maximize the expected number of users who are co-exposed to both campaigns. We show that the problem of maximizing co-exposure is NP-hard and its objective function is neither submodular nor supermodular. However, by exploiting a connection to a submodular function that acts as a lower bound to the objective, we are able to devise a greedy algorithm with provable approximation guarantee. We further provide a scalable instantiation of our approximation algorithm by introducing a novel extension to the notion of random reverse-reachable sets for efficiently estimating the expected co-exposure. We experimentally demonstrate the quality of our proposal on real-world social networks.

[UCLID-Net: Single View Reconstruction in Object Space](#)

- Benoit Guillard, Edoardo Remelli, Pascal Fua
- abstract@[open-review](#): Most state-of-the-art deep geometric learning single-view reconstruction approaches rely on encoder-decoder architectures that output either shape parametrizations or implicit representations. However, these representations rarely preserve the Euclidean structure of the 3D space objects exist in. In this paper, we show that building a geometry preserving 3-dimensional latent space helps the network concurrently learn global shape regularities and local reasoning in the object coordinate space and, as a result, boosts performance. We demonstrate both on ShapeNet synthetic images, which are often used for benchmarking purposes, and on real-world images that our approach outperforms state-of-the-art ones. Furthermore, the single-view pipeline naturally extends to multi-view reconstruction, which we also show.

[Reinforcement Learning for Control with Multiple Frequencies](#)

- Jongmin Lee, Byung-Jun Lee, Kee-Eung Kim
- abstract@[open-review](#): Many real-world sequential decision problems involve multiple action variables whose control frequencies are different, such that actions take their effects at different periods. While these problems can be formulated with the notion of multiple action persistences in factored-action MDP (FA-MDP), it is non-trivial to solve them efficiently since an action-persistent policy constructed from a stationary policy can be arbitrarily suboptimal, rendering solution methods for the standard FA-MDPs hardly applicable. In this paper, we formalize the problem of multiple control frequencies in RL and provide its efficient solution method. Our proposed method, Action-Persistent Policy Iteration (AP-PI), provides a theoretical guarantee on the convergence to an optimal solution while incurring only a factor of $\$|A|\$$ increase in time complexity during policy improvement step, compared to the standard policy iteration for FA-MDPs. Extending this result, we present Action-Persistent Actor-Critic (AP-AC), a scalable RL algorithm for high-dimensional control tasks. In the experiments, we demonstrate that AP-AC significantly outperforms the baselines on several continuous control tasks and a traffic control simulation, which highlights the effectiveness of our method that directly optimizes the periodic non-stationary policy for tasks with multiple control frequencies.

[Complex Dynamics in Simple Neural Networks: Understanding Gradient Flow in Phase Retrieval](#)

- Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, Pierfrancesco Urbani, Lenka Zdeborová
- abstract@[open-review](#): Despite the widespread use of gradient-based algorithms for optimising high-dimensional non-convex functions, understanding their ability of finding good minima instead of being trapped in spurious ones remains to a large extent an open problem. Here we focus on gradient flow dynamics for phase retrieval from random measurements. When the ratio of the number of measurements over the input dimension is small the dynamics remains trapped in spurious minima with large basins of attraction. We find analytically that above a critical ratio those critical points become unstable developing a negative direction toward the signal. By numerical experiments we show that in this regime the gradient flow algorithm is not trapped; it drifts away from the spurious critical points along the unstable direction and succeeds in finding the global minimum. Using tools from statistical physics we characterise this phenomenon, which is related to a BBP-type transition in the Hessian of the spurious minima.

[Neural Message Passing for Multi-Relational Ordered and Recursive Hypergraphs](#)

- Naganand Yadati
- abstract@[open-review](#): Message passing neural network (MPNN) has recently emerged as a successful framework by achieving state-of-the-art performances on many graph-based learning tasks. MPNN has also recently been extended to multi-relational graphs (each edge is labelled), and hypergraphs (each edge can connect any number of vertices). However, in real-world datasets involving text and knowledge, relationships are much more complex in which hyperedges can be multi-relational, recursive, and ordered. Such structures present several unique challenges because it is not clear how to adapt MPNN to variable-sized hyperedges in them. In this work, we first unify existing MPNNs on different structures into G-MPNN (Generalised MPNN) framework. Motivated by real-world datasets, we then propose a novel extension of the framework, MPNN-R (MPNN-Recursive) to handle recursively-structured data. Experimental results demonstrate the effectiveness of proposed G-MPNN and MPNN-R.

[A Unified View of Label Shift Estimation](#)

- Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, Zachary Lipton
- abstract@[open-review](#): Under label shift, the label distribution $p(y)$ might change but the class-conditional distributions $p(x|y)$ do not. There are two dominant approaches for estimating the label marginal. BBSE, a moment-matching approach based on confusion matrices, is provably consistent and

provides interpretable error bounds. However, a maximum likelihood estimation approach, which we call MLLS, dominates empirically. In this paper, we present a unified view of the two methods and the first theoretical characterization of MLLS. Our contributions include (i) consistency conditions for MLLS, which include calibration of the classifier and a confusion matrix invertibility condition that BBSE also requires; (ii) a unified framework, casting BBSE as roughly equivalent to MLLS for a particular choice of calibration method; and (iii) a decomposition of MLLS's finite-sample error into terms reflecting miscalibration and estimation error. Our analysis attributes BBSE's statistical inefficiency to a loss of information due to coarse calibration. Experiments on synthetic data, MNIST, and CIFAR10 support our findings.

[Optimal Private Median Estimation under Minimal Distributional Assumptions](#)

- Christos Tzamos, Emmanouil-Vasileios Vlatakis-Gkaragkounis, Ilias Zadik
- abstract@[open-review](#): We study the fundamental task of estimating the median of an underlying distribution from a finite number of samples, under pure differential privacy constraints. We focus on distributions satisfying the minimal assumption that they have a positive density at a small neighborhood around the median. In particular, the distribution is allowed to output unbounded values and is not required to have finite moments. We compute the exact, up-to-constant terms, statistical rate of estimation for the median by providing nearly-tight upper and lower bounds. Furthermore, we design a polynomial-time differentially private algorithm which provably achieves the optimal performance. At a technical level, our results leverage a Lipschitz Extension Lemma which allows us to design and analyze differentially private algorithms solely on appropriately defined ``typical'' instances of the samples.

[Breaking the Communication-Privacy-Accuracy Trilemma](#)

- Wei-Ning Chen, Peter Kairouz, Ayfer Ozgur
- abstract@[open-review](#): Two major challenges in distributed learning and estimation are 1) preserving the privacy of the local samples; and 2) communicating them efficiently to a central server, while achieving high accuracy for the end-to-end task. While there has been significant interest in addressing each of these challenges separately in the recent literature, treatments that simultaneously address both challenges are still largely missing. In this paper, we develop novel encoding and decoding mechanisms that simultaneously achieve optimal privacy and communication efficiency in various canonical settings. In particular, we consider the problems of mean estimation and frequency estimation under epsilon-local differential privacy and b-bit communication constraints. For mean estimation, we propose a scheme based on Kashinâ€™s representation and random sampling, with order-optimal estimation error under both constraints. For frequency estimation, we present a mechanism that leverages the recursive structure of Walsh-Hadamard matrices and achieves order-optimal estimation error for all privacy levels and communication budgets. As a by-product, we also construct a distribution estimation mechanism that is rate-optimal for all privacy regimes and communication constraints, extending recent work that is limited to $b = 1$ and $\epsilon = O(1)$. Our results demonstrate that intelligent encoding under joint privacy and communication constraints can yield a performance that matches the optimal accuracy achievable under either constraint alone.

[Audeo: Audio Generation for a Silent Performance Video](#)

- Kun Su, Xiulong Liu, Eli Shlizerman
- abstract@[open-review](#): We present a novel system that gets as an input, video frames of a musician playing the piano, and generates the music for that video. The generation of music from visual cues is a challenging problem and it is not clear whether it is an attainable goal at all. Our main aim in this work is to explore the plausibility of such a transformation and to identify cues and components able to carry the association of sounds with visual events. To achieve the transformation we built a full pipeline named 'Audeo' containing three components. We first translate the video frames of the keyboard and the musician hand movements into raw mechanical musical symbolic representation Piano-Roll (Roll) for each video frame which represents the keys pressed at each time step. We then adapt the Roll to be amenable for audio synthesis by including temporal correlations. This step turns out to be critical for meaningful audio generation. In the last step, we implement Midi synthesizers to generate realistic music. Audeo converts video to audio smoothly and clearly with only a few setup constraints. We evaluate Audeo on piano performance videos collected from YouTube and obtain that their generated music is of reasonable audio quality and can be successfully recognized with high precision by popular music identification software.

[Ode to an ODE](#)

- Krzysztof M. Choromanski, Jared Quincy Davis, Valerii Likhoshesterov, Xingyou Song, Jean-Jacques Slotine, Jacob Varley, Honglak Lee, Adrian Weller, Vikas Sindhwani
- abstract@[open-review](#): We present a new paradigm for Neural ODE algorithms, called ODEtoODE, where time-dependent parameters of the main flow evolve according to a matrix flow on the orthogonal group $O(d)$. This nested system of two flows, where the parameter-flow is constrained to lie on the compact manifold, provides stability and effectiveness of training and solves the gradient vanishing-explosion problem which is intrinsically related to training deep neural network architectures such as Neural ODEs. Consequently, it leads to better downstream models, as we show on the example of training reinforcement learning policies with evolution strategies, and in the supervised learning setting, by comparing with previous SOTA baselines. We provide strong convergence results for our proposed mechanism that are independent of the width of the network, supporting our empirical studies. Our results show an intriguing connection between the theory of deep neural networks and the field of matrix flows on compact manifolds.

[Self-Distillation Amplifies Regularization in Hilbert Space](#)

- Hossein Mobahi, Mehrdad Farajtabar, Peter Bartlett
- abstract@[open-review](#): Knowledge distillation introduced in the deep learning context is a method to transfer knowledge from one architecture to another. In particular, when the architectures are identical, this is called self-distillation. The idea is to feed in predictions of the trained model as new target values for retraining (and iterate this loop possibly a few times). It has been empirically observed that the self-distilled model often achieves higher accuracy on held out data. Why this happens, however, has been a mystery: the self-distillation dynamics does not receive any new information about the task and solely evolves by looping over training. To the best of our knowledge, there is no rigorous understanding of why this happens. This work provides the first theoretical analysis of self-distillation. We focus on fitting a nonlinear function to training data, where the model space is Hilbert space and fitting is subject to L2 regularization in this function space. We show that self-distillation iterations modify regularization by progressively limiting the number of basis functions that can be used to represent the solution. This implies (as we also verify empirically) that while a few rounds of self-distillation may reduce over-fitting, further rounds may lead to under-fitting and thus worse performance.

[Coupling-based Invertible Neural Networks Are Universal Diffeomorphism Approximators](#)

- Takeshi Teshima, Isao Ishikawa, Koichi Tojo, Kenta Oono, Masahiro Ikeda, Masashi Sugiyama
- abstract@[open-review](#): Invertible neural networks based on coupling flows (CF-INNs) have various machine learning applications such as image synthesis and representation learning. However, their desirable characteristics such as analytic invertibility come at the cost of restricting the functional forms. This poses a question on their representation power: are CF-INNs universal approximators for invertible functions? Without a universality, there could be a well-behaved invertible transformation that the CF-INN can never approximate, hence it would render the model class unreliable. We answer this question by showing a convenient criterion: a CF-INN is universal if its layers contain affine coupling and invertible linear functions as special cases. As its corollary, we can affirmatively resolve a previously unsolved problem: whether normalizing flow models based on affine coupling can be universal distributional approximators. In the course of proving the universality, we prove a general theorem to show the equivalence of the universality for certain diffeomorphism classes, a theoretical insight that is of interest by itself.

[Community detection using fast low-cardinality semidefinite programming](#)

- Po-Wei Wang, J. Zico Kolter
- abstract@[open-review](#): Modularity maximization has been a fundamental tool for understanding the community structure of a network, but the underlying optimization problem is nonconvex and NP-hard to solve. State-of-the-art algorithms like the Louvain or Leiden methods focus on different heuristics to help escape local optima, but they still depend on a greedy step that moves node assignment locally and is prone to getting trapped. In this paper, we propose a new class of low-cardinality algorithm that generalizes the local update to maximize a semidefinite relaxation derived from max-k-cut. This proposed algorithm is scalable, empirically achieves the global semidefinite optimality for small cases, and outperforms the state-of-the-art algorithms in real-world datasets with little additional time cost. From the algorithmic perspective, it also opens a new avenue for scaling-up semidefinite programming when the solutions are sparse instead of low-rank.

[Modeling Noisy Annotations for Crowd Counting](#)

- Jia Wan, Antoni Chan
- abstract@[open-review](#): The annotation noise in crowd counting is not modeled in traditional crowd counting algorithms based on crowd density maps. In this paper, we first model the annotation noise using a random variable with Gaussian distribution, and derive the pdf of the crowd density value for each spatial location in the image. We then approximate the joint distribution of the density values (i.e., the distribution of density maps) with a full covariance multivariate Gaussian density, and derive a low-rank approximate for tractable implementation. We use our loss function to train a crowd density map estimator and achieve state-of-the-art performance on three large-scale crowd counting datasets, which confirms its effectiveness. Examination of the predictions of the trained model shows that it can correctly predict the locations of people in spite of the noisy training data, which demonstrates the robustness of our loss function to annotation noise.

[An operator view of policy gradient methods](#)

- Dibya Ghosh, Marlos C. Machado, Nicolas Le Roux
- abstract@[open-review](#): We cast policy gradient methods as the repeated application of two operators: a policy improvement operator \mathcal{I} , which maps any policy π to a better one $\mathcal{I}\pi$, and a projection operator \mathcal{P} , which finds the best approximation of $\mathcal{I}\pi$ in the set of realizable policies. We use this framework to introduce operator-based versions of well-known policy gradient methods such as REINFORCE and PPO, which leads to a better understanding of their original counterparts. We also use the understanding we develop of the role of \mathcal{I} and \mathcal{P} to propose a new global lower bound of the expected return. This new perspective allows us to further bridge the gap between policy-based and value-based methods, showing how REINFORCE and the Bellman optimality operator, for example, can be seen as two sides of the same coin.

[Demystifying Contrastive Self-Supervised Learning: Invariances, Augmentations and Dataset Biases](#)

- Senthil Purushwalkam, Abhinav Gupta
- abstract@[open-review](#): Self-supervised representation learning approaches have recently surpassed their supervised learning counterparts on downstream tasks like object detection and image classification. Somewhat mysteriously the recent gains in performance come from training instance classification models, treating each image and its augmented versions as samples of a single class. In this work, we first present quantitative experiments to demystify these gains. We demonstrate that approaches like MOCO and PIRL learn occlusion-invariant representations. However, they fail to capture viewpoint and category instance invariance which are crucial components for object recognition. Second, we demonstrate that these approaches obtain further gains from access to a clean object-centric training dataset like Imagenet. Finally, we propose an approach to leverage unstructured videos to learn representations that possess higher viewpoint invariance. Our results show that the learned representations outperform MOCOv2 trained on the same data in terms of invariances encoded and the performance on downstream image classification and semantic segmentation tasks.

[Online MAP Inference of Determinantal Point Processes](#)

- Aditya Bhaskara, Amin Karbasi, Silvio Lattanzi, Morteza Zadimoghaddam
- abstract@[open-review](#): In this paper, we provide an efficient approximation algorithm for finding the most likelihood configuration (MAP) of size k for Determinantal Point Processes (DPP) in the online setting where the data points arrive in an arbitrary order and the algorithm cannot discard the selected elements from its local memory. Given a tolerance additive error η , our online algorithm achieves a $k^{O(k)}$ multiplicative approximation guarantee with an additive error η , using a memory footprint independent of the size of the data stream. We note that the exponential dependence on k in the approximation factor is unavoidable even in the offline setting. Our result readily implies a streaming algorithm with an improved memory bound compared to existing results.

[Video Object Segmentation with Adaptive Feature Bank and Uncertain-Region Refinement](#)

- Yongqing Liang, Xin Li, Navid Jafari, Jim Chen
- abstract@[open-review](#): This paper presents a new matching-based framework for semi-supervised video object segmentation (VOS). Recently, state-of-the-art VOS performance has been achieved by matching-based algorithms, in which feature banks are created to store features for region matching and classification. However, how to effectively organize information in the continuously growing feature bank remains under-explored, and this leads to an inefficient design of the bank. We introduced an adaptive feature bank update scheme to dynamically absorb new features and discard obsolete features. We also designed a new confidence loss and a fine-grained segmentation module to enhance the segmentation accuracy in uncertain regions. On public benchmarks, our algorithm outperforms existing state-of-the-arts.

[Inferring learning rules from animal decision-making](#)

- Zoe Ashwood, Nicholas A. Roy, Ji Hyun Bak, Jonathan W. Pillow
- abstract@[open-review](#): How do animals learn? This remains an elusive question in neuroscience. Whereas reinforcement learning often focuses on the design of algorithms that enable artificial agents to efficiently learn new tasks, here we develop a modeling framework to directly infer the empirical learning rules that animals use to acquire new behaviors. Our method efficiently infers the trial-to-trial changes in an animal's policy, and decomposes those changes into a learning component and a noise component. Specifically, this allows us to: (i) compare different learning rules and objective functions that an animal may be using to update its policy; (ii) estimate distinct learning rates for different parameters of an animal's policy; (iii) identify variations in learning across cohorts of animals; and (iv) uncover trial-to-trial changes that are not captured by normative learning rules. After validating our framework on simulated choice data, we applied our model to data from rats and mice learning perceptual decision-making tasks. We found that certain learning rules were far more capable of explaining trial-to-trial changes in an animal's policy. Whereas the average contribution of the conventional REINFORCE learning rule to the policy update for mice learning the International Brain Laboratory's task was just 30%, we found that adding baseline parameters allowed the learning rule to explain 92% of the animals' policy updates under our model. Intriguingly, the best-fitting learning rates and baseline values indicate that an animal's policy update, at each trial, does not occur in the direction that maximizes expected reward. Understanding how an animal transitions from chance-level to high-accuracy performance when learning a new task not only provides neuroscientists with insight into their animals, but also provides concrete examples of biological learning algorithms to the machine learning community.

[Input-Aware Dynamic Backdoor Attack](#)

- Tuan Anh Nguyen, Anh Tran
- abstract@[open-review](#): In recent years, neural backdoor attack has been considered to be a potential security threat to deep learning systems. Such systems, while achieving the state-of-the-art performance on clean data, perform abnormally on inputs with predefined triggers. Current backdoor techniques, however, rely on uniform trigger patterns, which are easily detected and mitigated by current defense methods. In this work, we propose a novel backdoor attack technique in which the triggers vary from input to input. To achieve this goal, we implement an input-aware trigger generator driven by diversity loss. A novel cross-trigger test is applied to enforce trigger nonreusability, making backdoor verification impossible. Experiments show that our method is efficient in various attack scenarios as well as multiple datasets. We further demonstrate that our backdoor can bypass the state of the art defense methods. An analysis with a famous neural network inspector again proves the stealthiness of the proposed attack. Our code is publicly available.

[How hard is to distinguish graphs with graph neural networks?](#)

- Andreas Loukas
- abstract@[open-review](#): A hallmark of graph neural networks is their ability to distinguish the isomorphism class of their inputs. This study derives hardness results for the classification variant of graph isomorphism in the message-passing model (MPNN). MPNN encompasses the majority of graph neural networks used today and is universal when nodes are given unique features. The analysis relies on the introduced measure of communication capacity. Capacity measures how much information the nodes of a network can exchange during the forward pass and depends on the depth, message-size, global state, and width of the architecture. It is shown that the capacity of MPNN needs to grow linearly with the number of nodes so that a network can distinguish trees and quadratically for general connected graphs. The derived bounds concern both worst- and average-case behavior and apply to networks with/without unique features and adaptive architecture---they are also up to two orders of magnitude tighter than those given by simpler arguments. An empirical study involving 12 graph classification tasks and 420 networks reveals strong alignment between actual performance and theoretical predictions.

[Minimax Regret of Switching-Constrained Online Convex Optimization: No Phase Transition](#)

- Lin Chen, Qian Yu, Hannah Lawrence, Amin Karbasi
- abstract@[open-review](#): We study the problem of switching-constrained online convex optimization (OCO), where the player has a limited number of opportunities to change her action. While the discrete analog of this online learning task has been studied extensively, previous work in the continuous setting has neither established the minimax rate nor algorithmically achieved it. In this paper, we show that $\$T\$$ -round switching-constrained OCO with fewer than $\$K\$$ switches has a minimax regret of $\$ \Theta(\frac{T}{\sqrt{K}}) \$$. In particular, it is at least $\$ \frac{T}{\sqrt{2K}} \$$ for one dimension and at least $\$ \frac{T}{\sqrt{K}} \$$ for higher dimensions. The lower bound in higher dimensions is attained by an orthogonal subspace argument. In one dimension, a novel adversarial strategy yields the lower bound of $\$O(\frac{T}{\sqrt{K}})\$, but a precise minimax analysis including constants is more involved. To establish the tighter one-dimensional result, we introduce the *fugal game* relaxation, whose minimax regret lower bounds that of switching-constrained OCO. We show that the minimax regret of the fugal game is at least $\$ \frac{T}{\sqrt{2K}} \$$ and thereby establish the optimal minimax lower bound in one dimension. To establish the dimension-independent upper bound, we next show that a mini-batching algorithm provides an $\$O(\frac{T}{\sqrt{K}})\$$ upper bound, and therefore conclude that the minimax regret of switching-constrained OCO is $\$ \Theta(\frac{T}{\sqrt{K}}) \$$ for any $\$K\$$. This is in sharp contrast to its discrete counterpart, the switching-constrained prediction-from-experts problem, which exhibits a phase transition in minimax regret between the low-switching and high-switching regimes.$

[Dual Manifold Adversarial Robustness: Defense against Lp and non-Lp Adversarial Attacks](#)

- Wei-An Lin, Chun Pong Lau, Alexander Levine, Rama Chellappa, Soheil Feizi
- abstract@[open-review](#): Adversarial training is a popular defense strategy against attack threat models with bounded L_p norms. However, it often degrades the model performance on normal images and more importantly, the defense does not generalize well to novel attacks. Given the success of deep generative models such as GANs and VAEs in characterizing the underlying manifold of images, we investigate whether or not the aforementioned deficiencies of adversarial training can be remedied by exploiting the underlying manifold information. To partially answer this question, we consider the scenario when the manifold information of the underlying data is available. We use a subset of ImageNet natural images where an approximate underlying manifold is learned using StyleGAN. We also construct an ``On-Manifold ImageNet'' (OM-ImageNet) dataset by projecting the ImageNet samples onto the learned manifold. For OM-ImageNet, the underlying manifold information is exact. Using OM-ImageNet, we first show that on-manifold adversarial training improves both standard accuracy and robustness to on-manifold attacks. However, since no out-of-manifold perturbations are realized, the defense can be broken by L_p adversarial attacks. We further propose Dual Manifold Adversarial Training (DMAT) where adversarial perturbations in both latent and image spaces are used in robustifying the model. Our DMAT improves performance on normal images, and achieves comparable robustness to the standard adversarial training against L_p attacks. In addition, we observe that models defended by DMAT achieve improved robustness against novel attacks which manipulate images by global color shifts or various types of image filtering. Interestingly, similar improvements are also achieved when the defended models are tested on (out-of-manifold) natural images. These results demonstrate the potential benefits of using manifold information in enhancing robustness of deep learning models against various types of novel adversarial attacks.

[Cross-Scale Internal Graph Neural Network for Image Super-Resolution](#)

- Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, Chen Change Loy
- abstract@[open-review](#): Non-local self-similarity in natural images has been well studied as an effective prior in image restoration. However, for single image super-resolution (SISR), most existing deep non-local methods (e.g., non-local neural networks) only exploit similar patches within the same scale of the low-resolution (LR) input image. Consequently, the restoration is limited to using the same-scale information while neglecting potential high-resolution (HR) cues from other scales. In this paper, we explore the cross-scale patch recurrence property of a natural image, i.e., similar patches tend to recur many times across different scales. This is achieved using a novel cross-scale internal graph neural network (IGNN). Specifically, we dynamically construct a cross-scale graph by searching k -nearest neighboring patches in the downsampled LR image for each query patch in the LR image. We then obtain the corresponding k HR neighboring patches in the LR image and aggregate them adaptively in accordance to the edge label of the constructed graph. In this way, the HR information can be passed from k HR neighboring patches to the LR query patch to help it recover more detailed textures. Besides, these internal image-specific LR/HR exemplars are also significant complements to the external information learned from the training dataset. Extensive experiments demonstrate the effectiveness of IGNN against the state-of-the-art SISR methods including existing non-local networks on standard benchmarks.

[Unsupervised Representation Learning by Invariance Propagation](#)

- Feng Wang, Huaping Liu, Di Guo, Sun Fuchun
- abstract@[open-review](#): Unsupervised learning methods based on contrastive learning have drawn increasing attention and achieved promising results. Most of them aim to learn representations invariant to instance-level variations, which are provided by different views of the same instance. In this paper, we propose Invariance Propagation to focus on learning representations invariant to category-level variations, which are provided by different instances from the same category. Our method recursively discovers semantically consistent samples residing in the same high-density regions in representation space. We demonstrate a hard sampling strategy to concentrate on maximizing the agreement between the anchor sample and its hard positive samples, which provide more intra-class variations to help capture more abstract invariance. As a result, with a ResNet-50 as the backbone, our method achieves

71.3% top-1 accuracy on ImageNet linear classification and 78.2% top-5 accuracy fine-tuning on only 1% labels, surpassing previous results. We also achieve state-of-the-art performance on other downstream tasks, including linear classification on Places205 and Pascal VOC, and transfer learning on small scale datasets.

[Restoring Negative Information in Few-Shot Object Detection](#)

- Yukuan Yang, Fangyun Wei, Miaojing Shi, Guoqi Li
- abstract@[open-review](#): Few-shot learning has recently emerged as a new challenge in the deep learning field: unlike conventional methods that train the deep neural networks (DNNs) with a large number of labeled data, it asks for the generalization of DNNs on new classes with few annotated samples. Recent advances in few-shot learning mainly focus on image classification while in this paper we focus on object detection. The initial explorations in few-shot object detection tend to simulate a classification scenario by using the positive proposals in images with respect to certain object class while discarding the negative proposals of that class. Negatives, especially hard negatives, however, are essential to the embedding space learning in few-shot object detection. In this paper, we restore the negative information in few-shot object detection by introducing a new negative- and positive-representative based metric learning framework and a new inference scheme with negative and positive representatives. We build our work on a recent few-shot pipeline RepMet with several new modules to encode negative information for both training and testing. Extensive experiments on ImageNet-LOC and PASCAL VOC show our method substantially improves the state-of-the-art few-shot object detection solutions. Our code is available at <https://github.com/yang-yk/NP-RepMet>.

[Do Adversarially Robust ImageNet Models Transfer Better?](#)

- Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, Aleksander Madry
- abstract@[open-review](#): Transfer learning is a widely-used paradigm in deep learning, where models pre-trained on standard datasets can be efficiently adapted to downstream tasks. Typically, better pre-trained models yield better transfer results, suggesting that initial accuracy is a key aspect of transfer learning performance. In this work, we identify another such aspect: we find that adversarially robust models, while less accurate, often perform better than their standard-trained counterparts when used for transfer learning. Specifically, we focus on adversarially robust ImageNet classifiers, and show that they yield improved accuracy on a standard suite of downstream classification tasks. Further analysis uncovers more differences between robust and standard models in the context of transfer learning. Our results are consistent with (and in fact, add to) recent hypotheses stating that robustness leads to improved feature representations. Code and models is available in the supplementary material.

[Robust Correction of Sampling Bias using Cumulative Distribution Functions](#)

- Bijan Mazaheri, Siddharth Jain, Jehoshua Bruck
- abstract@[open-review](#): Varying domains and biased datasets can lead to differences between the training and the target distributions, known as covariate shift. Current approaches for alleviating this often rely on estimating the ratio of training and target probability density functions. These techniques require parameter tuning and can be unstable across different datasets. We present a new method for handling covariate shift using the empirical cumulative distribution function estimates of the target distribution by a rigorous generalization of a recent idea proposed by Vapnik and Izmailov. Further, we show experimentally that our method is more robust in its predictions, is not reliant on parameter tuning and shows similar classification performance compared to the current state-of-the-art techniques on synthetic and real datasets.

[Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach](#)

- Alireza Fallah, Aryan Mokhtari, Asuman Ozdaglar
- abstract@[open-review](#): In Federated Learning, we aim to train models across multiple computing units (users), while users can only communicate with a common central server, without exchanging their data samples. This mechanism exploits the computational power of all users and allows users to obtain a richer model as their models are trained over a larger set of data points. However, this scheme only develops a common output for all the users, and, therefore, it does not adapt the model to each user. This is an important missing feature, especially given the heterogeneity of the underlying data distribution for various users. In this paper, we study a personalized variant of the federated learning in which our goal is to find an initial shared model that current or new users can easily adapt to their local dataset by performing one or a few steps of gradient descent with respect to their own data. This approach keeps all the benefits of the federated learning architecture, and, by structure, leads to a more personalized model for each user. We show this problem can be studied within the Model-Agnostic Meta-Learning (MAML) framework. Inspired by this connection, we study a personalized variant of the well-known Federated Averaging algorithm and evaluate its performance in terms of gradient norm for non-convex loss functions. Further, we characterize how this performance is affected by the closeness of underlying distributions of user data, measured in terms of distribution distances such as Total Variation and 1-Wasserstein metric.

[Pixel-Level Cycle Association: A New Perspective for Domain Adaptive Semantic Segmentation](#)

- Guoliang Kang, Yunchao Wei, Yi Yang, Yueling Zhuang, Alexander Hauptmann
- abstract@[open-review](#): Domain adaptive semantic segmentation aims to train a model performing satisfactory pixel-level predictions on the target with only out-of-domain (source) annotations. The conventional solution to this task is to minimize the discrepancy between source and target to enable effective knowledge transfer. Previous domain discrepancy minimization methods are mainly based on the adversarial training. They tend to consider the domain discrepancy globally, which ignore the pixel-wise relationships and are less discriminative. In this paper, we propose to build the pixel-level cycle association between source and target pixel pairs and contrastively strengthen their connections to diminish the domain gap and make the features more discriminative. To the best of our knowledge, this is a new perspective for tackling such a challenging task. Experiment results on two representative domain adaptation benchmarks, i.e. GTAV \$\rightarrow\$ Cityscapes and SYNTHIA \$\rightarrow\$ Cityscapes, verify the effectiveness of our proposed method and demonstrate that our method performs favorably against previous state-of-the-arts. Our method can be trained end-to-end in one stage and introduce no additional parameters, which is expected to serve as a general framework and help ease future research in domain adaptive semantic segmentation. Code is available at <https://github.com/kg1-prml/Pixel-Level-Cycle-Association>.

[Classification with Valid and Adaptive Coverage](#)

- Yaniv Romano, Matteo Sesia, Emmanuel Candès
- abstract@[open-review](#): Conformal inference, cross-validation+, and the jackknife+ are hold-out methods that can be combined with virtually any machine learning algorithm to construct prediction sets with guaranteed marginal coverage. In this paper, we develop specialized versions of these techniques for categorical and unordered response labels that, in addition to providing marginal coverage, are also fully adaptive to complex data distributions, in the sense that they perform favorably in terms of approximate conditional coverage compared to alternative methods. The heart of our contribution is a novel conformity score, which we explicitly demonstrate to be powerful and intuitive for classification problems, but whose underlying principle is potentially far more general. Experiments on synthetic and real data demonstrate the practical value of our theoretical guarantees, as well as the statistical advantages of the proposed methods over the existing alternatives.

[Learning Global Transparent Models consistent with Local Contrastive Explanations](#)

- Tejaswini Pedapati, Avinash Balakrishnan, Karthikeyan Shanmugam, Amit Dhurandhar
- abstract@[open-review](#): There is a rich and growing literature on producing local contrastive/counterfactual explanations for black-box models (e.g. neural networks). In these methods, for an input, an explanation is in the form of a contrast point differing in very few features from the original input and lying in a different class. Other works try to build globally interpretable models like decision trees and rule lists based on the data using actual labels or based on the black-box models predictions. Although these interpretable global models can be useful, they may not be consistent with local explanations from a specific black-box of choice. In this work, we explore the question: Can we produce a transparent global model that is simultaneously accurate and consistent with the local (contrastive) explanations of the black-box model? We introduce a local consistency metric that quantifies if the local explanations for the black-box model are also applicable to the proxy/surrogate globally transparent model. Based on a key insight we propose a novel method where we create custom boolean features from local contrastive explanations of the black-box model and then train a globally transparent model that has higher local consistency compared with other known strategies in addition to being accurate.

[Learning to Approximate a Bregman Divergence](#)

- Ali Siahkamari, XIDE XIA, Venkatesh Saligrama, David Castaño, Brian Kulis
- abstract@[open-review](#): Bregman divergences generalize measures such as the squared Euclidean distance and the KL divergence, and arise throughout many areas of machine learning. In this paper, we focus on the problem of approximating an arbitrary Bregman divergence from supervision, and we provide a well-principled approach to analyzing such approximations. We develop a formulation and algorithm for learning arbitrary Bregman divergences based on approximating their underlying convex generating function via a piecewise linear function. We provide theoretical approximation bounds using our parameterization and show that the generalization error $\mathcal{O}(p(m^{-1/2}))$ for metric learning using our framework matches the known generalization error in the strictly less general Mahalanobis metric learning setting. We further demonstrate empirically that our method performs well in comparison to existing metric learning methods, particularly for clustering and ranking problems.

[Diverse Image Captioning with Context-Object Split Latent Spaces](#)

- Shweta Mahajan, Stefan Roth
- abstract@[open-review](#): Diverse image captioning models aim to learn one-to-many mappings that are innate to cross-domain datasets, such as of images and texts. Current methods for this task are based on generative latent variable models, eg. VAEs with structured latent spaces. Yet, the amount of multimodality captured by prior work is limited to that of the paired training data -- the true diversity of the underlying generative process is not fully captured. To address this limitation, we leverage the contextual descriptions in the dataset that explain similar contexts in different visual scenes. To this end, we introduce a novel factorization of the latent space, termed context-object split, to model diversity in contextual descriptions across images and texts within the dataset. Our framework not only enables diverse captioning through context-based pseudo supervision, but extends this to images with novel objects and without paired captions in the training data. We evaluate our COS-CVAE approach on the standard COCO dataset and on the held-out COCO dataset consisting of images with novel objects, showing significant gains in accuracy and diversity.

[Learning Disentangled Representations of Videos with Missing Data](#)

- Armand Comas, Chi Zhang, Zlatan Feric, Octavia Camps, Rose Yu
- abstract@[open-review](#): Missing data poses significant challenges while learning representations of video sequences. We present Disentangled Imputed Video autoEncoder (DIVE), a deep generative model that imputes and predicts future video frames in the presence of missing data. Specifically, DIVE introduces a missingness latent variable, disentangles the hidden video representations into static and dynamic appearance, pose, and missingness factors for each object, while it imputes each object trajectory where data is missing. On a moving MNIST dataset with various missing scenarios, DIVE outperforms the state of the art baselines by a substantial margin. We also present comparisons on a real-world MOTSChallenge pedestrian dataset, which demonstrates the practical value of our method in a more realistic setting. Our code can be found in <https://github.com/Rose-STL-Lab/DIVE>.

[Natural Graph Networks](#)

- Pim de Haan, Taco S. Cohen, Max Welling
- abstract@[open-review](#): A key requirement for graph neural networks is that they must process a graph in a way that does not depend on how the graph is described. Traditionally this has been taken to mean that a graph network must be equivariant to node permutations. Here we show that instead of equivariance, the more general concept of naturality is sufficient for a graph network to be well-defined, opening up a larger class of graph networks. We define global and local natural graph networks, the latter of which are as scalable as conventional message passing graph neural networks while being more flexible. We give one practical instantiation of a natural network on graphs which uses an equivariant message network parameterization, yielding good performance on several benchmarks.

[Continual Learning with Node-Importance based Adaptive Group Sparse Regularization](#)

- Sangwon Jung, Hongjoon Ahn, Sungmin Cha, Taesup Moon
- abstract@[open-review](#): We propose a novel regularization-based continual learning method, dubbed as Adaptive Group Sparsity based Continual Learning (AGS-CL), using two group sparsity-based penalties. Our method selectively employs the two penalties when learning each neural network node based on its importance, which is adaptively updated after learning each task. By utilizing the proximal gradient descent method, the exact sparsity and freezing of the model is guaranteed during the learning process, and thus, the learner explicitly controls the model capacity. Furthermore, as a critical detail, we re-initialize the weights associated with unimportant nodes after learning each task in order to facilitate efficient learning and prevent the negative transfer. Throughout the extensive experimental results, we show that our AGS-CL uses orders of magnitude less memory space for storing the regularization parameters, and it significantly outperforms several state-of-the-art baselines on representative benchmarks for both supervised and reinforcement learning.

[Towards Crowdsourced Training of Large Neural Networks using Decentralized Mixture-of-Experts](#)

- Max Ryabinin, Anton Gusev
- abstract@[open-review](#): Many recent breakthroughs in deep learning were achieved by training increasingly larger models on massive datasets. However, training such models can be prohibitively expensive. For instance, the cluster used to train GPT-3 costs over \$250 million. As a result, most researchers cannot afford to train state of the art models and contribute to their development. Hypothetically, a researcher could crowdsource the training of large neural networks with thousands of regular PCs provided by volunteers. The raw computing power of a hundred thousand \$2500 desktops dwarfs that of a \$250M server pod, but one cannot utilize that power efficiently with conventional distributed training methods. In this work, we propose Learning@home: a novel neural network training paradigm designed to handle large amounts of poorly connected participants. We analyze the performance, reliability, and architectural constraints of this paradigm and compare it against existing distributed training techniques.

[Bidirectional Convolutional Poisson Gamma Dynamical Systems](#)

- wenchochen, Chaojie Wang, Bo Chen, Yicheng Liu, Hao Zhang, Mingyuan Zhou

- abstract@[open-review](#): Incorporating the natural document-sentence-word structure into hierarchical Bayesian modeling, we propose convolutional Poisson gamma dynamical systems (PGDS) that introduce not only word-level probabilistic convolutions, but also sentence-level stochastic temporal transitions. With word-level convolutions capturing phrase-level topics and sentence-level transitions capturing how the topic usages evolve over consecutive sentences, we aggregate the topic proportions of all sentences of a document as its feature representation. To consider not only forward but also backward sentence-level information transmissions, we further develop a bidirectional convolutional PGDS to incorporate the full contextual information to represent each sentence. For efficient inference, we construct a convolutional-recurrent inference network, which provides both sentence-level and document-level representations, and introduce a hybrid Bayesian inference scheme combining stochastic-gradient MCMC and amortized variational inference. Experimental results on a variety of document corpora demonstrate that the proposed models can extract expressive multi-level latent representations, including interpretable phrase-level topics and sentence-level temporal transitions as well as discriminative document-level features, achieving state-of-the-art document categorization performance while being memory and computation efficient.

[Deep Reinforcement and InfoMax Learning](#)

- Bogdan Mazoure, Remi Tachet des Combes, Thang Long Doan, Philip Bachman, R Devon Hjelm
- abstract@[open-review](#): We posit that a reinforcement learning (RL) agent will perform better when it uses representations that are better at predicting the future, particularly in terms of few-shot learning and domain adaptation. To test that hypothesis, we introduce an objective based on Deep InfoMax (DIM) which trains the agent to predict the future by maximizing the mutual information between its internal representation of successive timesteps. We provide an intuitive analysis of the convergence properties of our approach from the perspective of Markov chain mixing times, and argue that convergence of the lower bound on mutual information is related to the inverse absolute spectral gap of the transition model. We test our approach in several synthetic settings, where it successfully learns representations that are predictive of the future. Finally, we augment C51, a strong distributional RL agent, with our temporal DIM objective and demonstrate on a continual learning task (inspired by Ms.~PacMan) and on the recently introduced Procgen environment that our approach improves performance, which supports our core hypothesis.

[On ranking via sorting by estimated expected utility](#)

- Clement Calauzenes, Nicolas Usunier
- abstract@[open-review](#): Ranking and selection tasks appear in different contexts with specific desiderata, such as the maximization of average relevance on the top of the list, the requirement of diverse rankings, or, relatedly, the focus on providing at least one relevant item to as many users as possible. This paper addresses the question of which of these tasks are asymptotically solved by sorting by decreasing order of expected utility, for some suitable notion of utility, or, equivalently, \emph{when is square loss regression consistent for ranking \emph{via} score-and-sort?}. We provide an answer to this question in the form of a structural characterization of ranking losses for which a suitable regression is consistent. This result has two fundamental corollaries. First, whenever there exists a consistent approach based on convex risk minimization, there also is a consistent approach based on regression. Second, when regression is not consistent, there are data distributions for which consistent surrogate approaches necessarily have non-trivial local minima, and optimal scoring function are necessarily discontinuous, even when the underlying data distribution is regular. In addition to providing a better understanding of surrogate approaches for ranking, these results illustrate the intrinsic difficulty of solving general ranking problems with the score-and-sort approach.

[Distribution-free binary classification: prediction sets, confidence intervals and calibration](#)

- Chirag Gupta, Aleksandr Podkopaev, Aaditya Ramdas
- abstract@[open-review](#): We study three notions of uncertainty quantification---calibration, confidence intervals and prediction sets---for binary classification in the distribution-free setting, that is without making any distributional assumptions on the data. With a focus towards calibration, we establish a 'tripod' of theorems that connect these three notions for score-based classifiers. A direct implication is that distribution-free calibration is only possible, even asymptotically, using a scoring function whose level sets partition the feature space into at most countably many sets. Parametric calibration schemes such as variants of Platt scaling do not satisfy this requirement, while nonparametric schemes based on binning do. To close the loop, we derive distribution-free confidence intervals for binned probabilities for both fixed-width and uniform-mass binning. As a consequence of our 'tripod' theorems, these confidence intervals for binned probabilities lead to distribution-free calibration. We also derive extensions to settings with streaming data and covariate shift.

[Closing the Dequantization Gap: PixelCNN as a Single-Layer Flow](#)

- Didrik Nielsen, Ole Winther
- abstract@[open-review](#): Flow models have recently made great progress at modeling ordinal discrete data such as images and audio. Due to the continuous nature of flow models, dequantization is typically applied when using them for such discrete data, resulting in lower bound estimates of the likelihood. In this paper, we introduce subset flows, a class of flows that can tractably transform finite volumes and thus allow exact computation of likelihoods for discrete data. Based on subset flows, we identify ordinal discrete autoregressive models, including WaveNets, PixelCNNs and Transformers, as single-layer flows. We use the flow formulation to compare models trained and evaluated with either the exact likelihood or its dequantization lower bound. Finally, we study multilayer flows composed of PixelCNNs and non-autoregressive coupling layers and demonstrate state-of-the-art results on CIFAR-10 for flow models trained with dequantization.

[Sequence to Multi-Sequence Learning via Conditional Chain Mapping for Mixture Signals](#)

- Jing Shi, Xuankai Chang, Pengcheng Guo, Shinji Watanabe, Yusuke Fujita, Jiaming Xu, Bo Xu, Lei Xie
- abstract@[open-review](#): Neural sequence-to-sequence models are well established for applications which can be cast as mapping a single input sequence into a single output sequence. In this work, we focus on one-to-many sequence transduction problems, such as extracting multiple sequential sources from a mixture sequence. We extend the standard sequence-to-sequence model to a conditional multi-sequence model, which explicitly models the relevance between multiple output sequences with the probabilistic chain rule. Based on this extension, our model can conditionally infer output sequences one-by-one by making use of both input and previously-estimated contextual output sequences. This model additionally has a simple and efficient stop criterion for the end of the transduction, making it able to infer the variable number of output sequences. We take speech data as a primary test field to evaluate our methods since the observed speech data is often composed of multiple sources due to the nature of the superposition principle of sound waves. Experiments on several different tasks including speech separation and multi-speaker speech recognition show that our conditional multi-sequence models lead to consistent improvements over the conventional non-conditional models.

[Variance reduction for Random Coordinate Descent-Langevin Monte Carlo](#)

- ZHIYAN DING, Qin Li
- abstract@[open-review](#): Sampling from a log-concave distribution function is one core problem that has wide applications in Bayesian statistics and machine learning. While most gradient free methods have slow convergence rate, the Langevin Monte Carlo (LMC) that provides fast convergence requires the computation of gradients. In practice one uses finite-differencing approximations as surrogates, and the method is expensive in high-dimensions. A natural strategy to reduce computational cost in each iteration is to utilize random gradient approximations, such as random coordinate descent (RCD) or simultaneous perturbation stochastic approximation (SPSA). We show by a counterexample that blindly applying RCD does not achieve the goal in the most general setting. The high variance induced by the randomness means a larger number of iterations are needed, and this balances out

the saving in each iteration. We then introduce a new variance reduction approach, termed Randomized Coordinates Averaging Descent (RCAD), and incorporate it with both overdamped and underdamped LMC. The methods are termed RCAD-O-LMC and RCAD-U-LMC respectively. The methods still sit in the random gradient approximation framework, and thus the computational cost in each iteration is low. However, by employing RCAD, the variance is reduced, so the methods converge within the same number of iterations as the classical overdamped and underdamped LMC. This leads to a computational saving overall.

[Language as a Cognitive Tool to Imagine Goals in Curiosity Driven Exploration](#)

- Cédric Colas, Tristan Karch, Nicolas Lair, Jean-Michel Dussoux, Clément Moulin-Frier, Peter Dominey, Pierre-Yves Oudeyer
- abstract@[open-review](#): Developmental machine learning studies how artificial agents can model the way children learn open-ended repertoires of skills. Such agents need to create and represent goals, select which ones to pursue and learn to achieve them. Recent approaches have considered goal spaces that were either fixed and hand-defined or learned using generative models of states. This limited agents to sample goals within the distribution of known effects. We argue that the ability to imagine out-of-distribution goals is key to enable creative discoveries and open-ended learning. Children do so by leveraging the compositionality of language as a tool to imagine descriptions of outcomes they never experienced before, targeting them as goals during play. We introduce IMAGINE, an intrinsically motivated deep reinforcement learning architecture that models this ability. Such imaginative agents, like children, benefit from the guidance of a social peer who provides language descriptions. To take advantage of goal imagination, agents must be able to leverage these descriptions to interpret their imagined out-of-distribution goals. This generalization is made possible by modularity: a decomposition between learned goal-achievement reward function and policy relying on deep sets, gated attention and object-centered representations. We introduce the Playground environment and study how this form of goal imagination improves generalization and exploration over agents lacking this capacity. In addition, we identify the properties of goal imagination that enable these results and study the impacts of modularity and social interactions.

[All Word Embeddings from One Embedding](#)

- Sho Takase, Sosuke Kobayashi
- abstract@[open-review](#): In neural network-based models for natural language processing (NLP), the largest part of the parameters often consists of word embeddings. Conventional models prepare a large embedding matrix whose size depends on the vocabulary size. Therefore, storing these models in memory and disk storage is costly. In this study, to reduce the total number of parameters, the embeddings for all words are represented by transforming a shared embedding. The proposed method, ALONE (all word embeddings from one), constructs the embedding of a word by modifying the shared embedding with a filter vector, which is word-specific but non-trainable. Then, we input the constructed embedding into a feed-forward neural network to increase its expressiveness. Naively, the filter vectors occupy the same memory size as the conventional embedding matrix, which depends on the vocabulary size. To solve this issue, we also introduce a memory-efficient filter construction approach. We indicate our ALONE can be used as word representation sufficiently through an experiment on the reconstruction of pre-trained word embeddings. In addition, we also conduct experiments on NLP application tasks: machine translation and summarization. We combined ALONE with the current state-of-the-art encoder-decoder model, the Transformer [36], and achieved comparable scores on WMT 2014 English-to-German translation and DUC 2004 very short summarization with less parameters.

[Primal Dual Interpretation of the Proximal Stochastic Gradient Langevin Algorithm](#)

- Adil Salim, Peter Richtarik
- abstract@[open-review](#): We consider the task of sampling with respect to a log concave probability distribution. The potential of the target distribution is assumed to be composite, i.e., written as the sum of a smooth convex term, and a nonsmooth convex term possibly taking infinite values. The target distribution can be seen as a minimizer of the Kullback-Leibler divergence defined on the Wasserstein space (i.e., the space of probability measures). In the first part of this paper, we establish a strong duality result for this minimization problem. In the second part of this paper, we use the duality gap arising from the first part to study the complexity of the Proximal Stochastic Gradient Langevin Algorithm (PSGLA), which can be seen as a generalization of the Projected Langevin Algorithm. Our approach relies on viewing PSGLA as a primal dual algorithm and covers many cases where the target distribution is not fully supported. In particular, we show that if the potential is strongly convex, the complexity of PSGLA is $\mathcal{O}(1/\epsilon^2)$ in terms of the 2-Wasserstein distance. In contrast, the complexity of the Projected Langevin Algorithm is $\mathcal{O}(1/\epsilon^{12})$ in terms of total variation when the potential is convex.

[How to Characterize The Landscape of Overparameterized Convolutional Neural Networks](#)

- Yihong Gu, Weizhong Zhang, Cong Fang, Jason D. Lee, Tong Zhang
- abstract@[open-review](#): For many initialization schemes, parameters of two randomly initialized deep neural networks (DNNs) can be quite different, but feature distributions of the hidden nodes are similar at each layer. With the help of a new technique called {it neural network grafting}, we demonstrate that even during the entire training process, feature distributions of differently initialized networks remain similar at each layer. In this paper, we present an explanation of this phenomenon. Specifically, we consider the loss landscape of an overparameterized convolutional neural network (CNN) in the continuous limit, where the numbers of channels/hidden nodes in the hidden layers go to infinity. Although the landscape of the overparameterized CNN is still non-convex with respect to the trainable parameters, we show that very surprisingly, it can be reformulated as a convex function with respect to the feature distributions in the hidden layers. Therefore by reparameterizing neural networks in terms of feature distributions, we obtain a much simpler characterization of the landscape of overparameterized CNNs. We further argue that training with respect to network parameters leads to a fixed trajectory in the feature distributions.

[On the Tightness of Semidefinite Relaxations for Certifying Robustness to Adversarial Examples](#)

- Richard Zhang
- abstract@[open-review](#): The robustness of a neural network to adversarial examples can be provably certified by solving a convex relaxation. If the relaxation is loose, however, then the resulting certificate can be too conservative to be practically useful. Recently, a less conservative robustness certificate was proposed, based on a semidefinite programming (SDP) relaxation of the ReLU activation function. In this paper, we describe a geometric technique that determines whether this SDP certificate is exact, meaning whether it provides both a lower-bound on the size of the smallest adversarial perturbation, as well as a globally optimal perturbation that attains the lower-bound. Concretely, we show, for a least-squares restriction of the usual adversarial attack problem, that the SDP relaxation amounts to the nonconvex projection of a point onto a hyperbola. The resulting SDP certificate is exact if and only if the projection of the point lies on the major axis of the hyperbola. Using this geometric technique, we prove that the certificate is exact over a single hidden layer under mild assumptions, and explain why it is usually conservative for several hidden layers. We experimentally confirm our theoretical insights using a general-purpose interior-point method and a custom rank-2 Burer-Monteiro algorithm.

[Submodular Meta-Learning](#)

- Arman Adibi, Aryan Mokhtari, Hamed Hassani
- abstract@[open-review](#): In this paper, we introduce a discrete variant of the Meta-learning framework. Meta-learning aims at exploiting prior experience and data to improve performance on future tasks. By now, there exist numerous formulations for Meta-learning in the continuous domain. Notably, the Model-Agnostic Meta-Learning (MAML) formulation views each task as a continuous optimization problem and based on prior data learns a suitable initialization that can be adapted to new, unseen tasks after a few simple gradient updates. Motivated by this terminology, we propose a novel Meta-learning framework in the discrete domain where each task is equivalent to maximizing a set function under a cardinality constraint. Our approach aims at

using prior data, i.e., previously visited tasks, to train a proper initial solution set that can be quickly adapted to a new task at a relatively low computational cost. This approach leads to (i) a personalized solution for each task, and (ii) significantly reduced computational cost at test time compared to the case where the solution is fully optimized once the new task is revealed. The training procedure is performed by solving a challenging discrete optimization problem for which we present deterministic and randomized algorithms. In the case where the tasks are monotone and submodular, we show strong theoretical guarantees for our proposed methods even though the training objective may not be submodular. We also demonstrate the effectiveness of our framework on two real-world problem instances where we observe that our methods lead to a significant reduction in computational complexity in solving the new tasks while incurring a small performance loss compared to when the tasks are fully optimized.

[Rethinking Pre-training and Self-training](#)

- Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, Quoc Le
- abstract@[open-review](#): Pre-training is a dominant paradigm in computer vision. For example, supervised ImageNet pre-training is commonly used to initialize the backbones of object detection and segmentation models. He et al., however, show a striking result that ImageNet pre-training has limited impact on COCO object detection. Here we investigate self-training as another method to utilize additional data on the same setup and contrast it against ImageNet pre-training. Our study reveals the generality and flexibility of self-training with three additional insights: 1) stronger data augmentation and more labeled data further diminish the value of pre-training, 2) unlike pre-training, self-training is always helpful when using stronger data augmentation, in both low-data and high-data regimes, and 3) in the case that pre-training is helpful, self-training improves upon pre-training. For example, on the COCO object detection dataset, pre-training benefits when we use one fifth of the labeled data, and hurts accuracy when we use all labeled data. Self-training, on the other hand, shows positive improvements from +1.3 to +3.4AP across all dataset sizes. In other words, self-training works well exactly on the same setup that pre-training does not work (using ImageNet to help COCO). On the PASCAL segmentation dataset, which is a much smaller dataset than COCO, though pre-training does help significantly, self-training improves upon the pre-trained model. On COCO object detection, we achieve 53.8AP, an improvement of +1.7AP over the strongest SpineNet model. On PASCAL segmentation, we achieve 90.5mIOU, an improvement of +1.5mIOU over the previous state-of-the-art result by DeepLabv3+.

[Unsupervised Sound Separation Using Mixture Invariant Training](#)

- Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron Weiss, Kevin Wilson, John Hershey
- abstract@[open-review](#): In recent years, rapid progress has been made on the problem of single-channel sound separation using supervised training of deep neural networks. In such supervised approaches, a model is trained to predict the component sources from synthetic mixtures created by adding up isolated ground-truth sources. Reliance on this synthetic training data is problematic because good performance depends upon the degree of match between the training data and real-world audio, especially in terms of the acoustic conditions and distribution of sources. The acoustic properties can be challenging to accurately simulate, and the distribution of sound types may be hard to replicate. In this paper, we propose a completely unsupervised method, mixture invariant training (MixIT), that requires only single-channel acoustic mixtures. In MixIT, training examples are constructed by mixing together existing mixtures, and the model separates them into a variable number of latent sources, such that the separated sources can be remixed to approximate the original mixtures. We show that MixIT can achieve competitive performance compared to supervised methods on speech separation. Using MixIT in a semi-supervised learning setting enables unsupervised domain adaptation and learning from large amounts of real-world data without ground-truth source waveforms. In particular, we significantly improve reverberant speech separation performance by incorporating reverberant mixtures, train a speech enhancement system from noisy mixtures, and improve universal sound separation by incorporating a large amount of in-the-wild data.

[Adaptive Discretization for Model-Based Reinforcement Learning](#)

- Sean Sinclair, Tianyu Wang, Gauri Jain, Siddhartha Banerjee, Christina Yu
- abstract@[open-review](#): We introduce the technique of adaptive discretization to design an efficient model-based episodic reinforcement learning algorithm in large (potentially continuous) state-action spaces. Our algorithm is based on optimistic one-step value iteration extended to maintain an adaptive discretization of the space. From a theoretical perspective we provide worst-case regret bounds for our algorithm which are competitive compared to the state-of-the-art model-based algorithms. Moreover, our bounds are obtained via a modular proof technique which can potentially extend to incorporate additional structure on the problem. From an implementation standpoint, our algorithm has much lower storage and computational requirements due to maintaining a more efficient partition of the state and action spaces. We illustrate this via experiments on several canonical control problems, which shows that our algorithm empirically performs significantly better than fixed discretization in terms of both faster convergence and lower memory usage. Interestingly, we observe empirically that while fixed discretization model-based algorithms vastly outperform their model-free counterparts, the two achieve comparable performance with adaptive discretization.

[CodeCMR: Cross-Modal Retrieval For Function-Level Binary Source Code Matching](#)

- Zeping Yu, Wenxin Zheng, Jiaqi Wang, Qiyi Tang, Sen Nie, Shi Wu
- abstract@[open-review](#): Binary source code matching, especially on function-level, has a critical role in the field of computer security. Given binary code only, finding the corresponding source code improves the accuracy and efficiency in reverse engineering. Given source code only, related binary code retrieval contributes to known vulnerabilities confirmation. However, due to the vast difference between source and binary code, few studies have investigated binary source code matching. Previously published studies focus on code literals extraction such as strings and integers, then utilize traditional matching algorithms such as the Hungarian algorithm for code matching. Nevertheless, these methods have limitations on function-level, because they ignore the potential semantic features of code and a lot of code lacks sufficient code literals. Also, these methods indicate a need for expert experience for useful feature identification and feature engineering, which is timeconsuming. This paper proposes an end-to-end cross-modal retrieval network for binary source code matching, which achieves higher accuracy and requires less expert experience. We adopt Deep Pyramid Convolutional Neural Network (DPCNN) for source code feature extraction and Graph Neural Network (GNN) for binary code feature extraction. We also exploit neural network-based models to capture code literals, including strings and integers. Furthermore, we implement "norm weighted sampling" for negative sampling. We evaluate our model on two datasets, where it outperforms other methods significantly.

[On Warm-Starting Neural Network Training](#)

- Jordan Ash, Ryan P. Adams
- abstract@[open-review](#): In many real-world deployments of machine learning systems, data arrive piecemeal. These learning scenarios may be passive, where data arrive incrementally due to structural properties of the problem (e.g., daily financial data) or active, where samples are selected according to a measure of their quality (e.g., experimental design). In both of these cases, we are building a sequence of models that incorporate an increasing amount of data. We would like each of these models in the sequence to be performant and take advantage of all the data that are available to that point. Conventional intuition suggests that when solving a sequence of related optimization problems of this form, it should be possible to initialize using the solution of the previous iterate---to ``warm start'' the optimization rather than initialize from scratch---and see reductions in wall-clock time. However, in practice this warm-starting seems to yield poorer generalization performance than models that have fresh random initializations, even though the final training losses are similar. While it appears that some hyperparameter settings allow a practitioner to close this generalization gap, they seem to only do so in regimes that damage the wall-clock gains of the warm start. Nevertheless, it is highly desirable to be able to warm-start neural network training, as it would dramatically reduce the resource usage associated with the construction of performant deep learning systems. In this work, we take a closer look at this empirical phenomenon and try to understand when and how it occurs. We also provide a surprisingly simple trick that overcomes this pathology in several important situations, and present experiments that elucidate some of its properties.

DAGs with No Fears: A Closer Look at Continuous Optimization for Learning Bayesian Networks

- Dennis Wei, Tian Gao, Yue Yu
- abstract@[open-review](#): This paper re-examines a continuous optimization framework dubbed NOTEARS for learning Bayesian networks. We first generalize existing algebraic characterizations of acyclicity to a class of matrix polynomials. Next, focusing on a one-parameter-per-edge setting, it is shown that the Karush-Kuhn-Tucker (KKT) optimality conditions for the NOTEARS formulation cannot be satisfied except in a trivial case, which explains a behavior of the associated algorithm. We then derive the KKT conditions for an equivalent reformulation, show that they are indeed necessary, and relate them to explicit constraints that certain edges be absent from the graph. If the score function is convex, these KKT conditions are also sufficient for local minimality despite the non-convexity of the constraint. Informed by the KKT conditions, a local search post-processing algorithm is proposed and shown to substantially and universally improve the structural Hamming distance of all tested algorithms, typically by a factor of 2 or more. Some combinations with local search are both more accurate and more efficient than the original NOTEARS.

OOD-MAML: Meta-Learning for Few-Shot Out-of-Distribution Detection and Classification

- Taewon Jeong, Heeyoung Kim
- abstract@[open-review](#): We propose a few-shot learning method for detecting out-of-distribution (OOD) samples from classes that are unseen during training while classifying samples from seen classes using only a few labeled examples. For detecting unseen classes while generalizing to new samples of known classes, we synthesize fake samples, i.e., OOD samples, but that resemble in-distribution samples, and use them along with real samples. Our approach is based on an extension of model-agnostic meta learning (MAML) and is denoted as OOD-MAML, which not only learns a model initialization but also the initial fake samples across tasks. The learned initial fake samples can be used to quickly adapt to new tasks to form task-specific fake samples with only one or a few gradient update steps using MAML. For testing, OOD-MAML converts a K-shot N-way classification task into N sub-tasks of K-shot OOD detection with respect to each class. The joint analysis of N sub-tasks facilitates simultaneous classification and OOD detection and, furthermore, offers an advantage, in that it does not require re-training when the number of classes for a test task differs from that for training tasks; it is sufficient to simply assume as many sub-tasks as the number of classes for the test task. We also demonstrate the effective performance of OOD-MAML over benchmark datasets.

An Imitation from Observation Approach to Transfer Learning with Dynamics Mismatch

- Siddharth Desai, Ishan Durugkar, Haresh Karnan, Garrett Warnell, Josiah Hanna, Peter Stone
- abstract@[open-review](#): We examine the problem of transferring a policy learned in a source environment to a target environment with different dynamics, particularly in the case where it is critical to reduce the amount of interaction with the target environment during learning. This problem is particularly important in sim-to-real transfer because simulators inevitably model real-world dynamics imperfectly. In this paper, we show that one existing solution to this transfer problem-- grounded action transformation --is closely related to the problem of imitation from observation (IfO): learning behaviors that mimic the observations of behavior demonstrations. After establishing this relationship, we hypothesize that recent state-of-the-art approaches from the IfO literature can be effectively repurposed for grounded transfer learning. To validate our hypothesis we derive a new algorithm -- generative adversarial reinforced action transformation (GARAT) -- based on adversarial imitation from observation techniques. We run experiments in several domains with mismatched dynamics, and find that agents trained with GARAT achieve higher returns in the target environment compared to existing black-box transfer methods.

Learning About Objects by Learning to Interact with Them

- Martin Lohmann, Jordi Salvador, Aniruddha Kembhavi, Roozbeh Mottaghi
- abstract@[open-review](#): Much of the remarkable progress in computer vision has been focused around fully supervised learning mechanisms relying on highly curated datasets for a variety of tasks. In contrast, humans often learn about their world with little to no external supervision. Taking inspiration from infants learning from their environment through play and interaction, we present a computational framework to discover objects and learn their physical properties along this paradigm of Learning from Interaction. Our agent, when placed within the near photo-realistic and physics-enabled AI2-THOR environment, interacts with its world and learns about objects, their geometric extents and relative masses, without any external guidance. Our experiments reveal that this agent learns efficiently and effectively; not just for objects it has interacted with before, but also for novel instances from seen categories as well as novel object categories.

Learning discrete distributions with infinite support

- Doron Cohen, Aryeh Kontorovich, Georrey Wolfer
- abstract@[open-review](#): We present a novel approach to estimating discrete distributions with (potentially) infinite support in the total variation metric. In a departure from the established paradigm, we make no structural assumptions whatsoever on the sampling distribution. In such a setting, distribution-free risk bounds are impossible, and the best one could hope for is a fully empirical data-dependent bound. We derive precisely such bounds, and demonstrate that these are, in a well-defined sense, the best possible. Our main discovery is that the half-norm of the empirical distribution provides tight upper and lower estimates on the empirical risk. Furthermore, this quantity decays at a nearly optimal rate as a function of the true distribution. The optimality follows from a minimax result, of possible independent interest. Additional structural results are provided, including an exact Rademacher complexity calculation and apparently a first connection between the total variation risk and the missing mass.

Dissecting Neural ODEs

- Stefano Massaroli, Michael Poli, Jinkyoo Park, Atsushi Yamashita, Hajime Asama
- abstract@[open-review](#): Continuous deep learning architectures have recently re-emerged as Neural Ordinary Differential Equations (Neural ODEs). This infinite-depth approach theoretically bridges the gap between deep learning and dynamical systems, offering a novel perspective. However, deciphering the inner working of these models is still an open challenge, as most applications apply them as generic black-box modules. In this work we ``open the box'', further developing the continuous-depth formulation with the aim of clarifying the influence of several design choices on the underlying dynamics.

Teaching a GAN What Not to Learn

- Siddarth Asokan, Chandra Seelamantula
- abstract@[open-review](#): Generative adversarial networks (GANs) were originally envisioned as unsupervised generative models that learn to follow a target distribution. Variants such as conditional GANs, auxiliary-classifier GANs (ACGANs) project GANs on to supervised and semi-supervised learning frameworks by providing labelled data and using multi-class discriminators. In this paper, we approach the supervised GAN problem from a different perspective, one that is motivated by the philosophy of the famous Persian poet Rumi who said, "The art of knowing is knowing what to ignore." In the GAN framework, we not only provide the GAN positive data that it must learn to model, but also present it with so-called negative samples that it must learn to avoid -- we call this "The Rumi Framework." This formulation allows the discriminator to represent the underlying target distribution better by learning to penalize generated samples that are undesirable -- we show that this capability accelerates the learning process of the generator. We present a reformulation of the standard GAN (SGAN) and least-squares GAN (LSGAN) within the Rumi setting. The advantage of the reformulation is demonstrated by means of experiments conducted on MNIST, Fashion MNIST, CelebA, and CIFAR-10 datasets. Finally, we consider an application of the

proposed formulation to address the important problem of learning an under-represented class in an unbalanced dataset. The Rumi approach results in substantially lower FID scores than the standard GAN frameworks while possessing better generalization capability.

[Counterfactual Data Augmentation using Locally Factored Dynamics](#)

- Silviu Pitis, Elliot Creager, Animesh Garg
- abstract@[open-review](#): Many dynamic processes, including common scenarios in robotic control and reinforcement learning (RL), involve a set of interacting subprocesses. Though the subprocesses are not independent, their interactions are often sparse, and the dynamics at any given time step can often be decomposed into locally independent causal mechanisms. Such local causal structures can be leveraged to improve the sample efficiency of sequence prediction and off-policy reinforcement learning. We formalize this by introducing local causal models (LCMs), which are induced from a global causal model by conditioning on a subset of the state space. We propose an approach to inferring these structures given an object-oriented state representation, as well as a novel algorithm for Counterfactual Data Augmentation (CoDA). CoDA uses local structures and an experience replay to generate counterfactual experiences that are causally valid in the global model. We find that CoDA significantly improves the performance of RL agents in locally factored tasks, including the batch-constrained and goal-conditioned settings. Code available at <https://github.com/spitis/mrl>.

[Rethinking Learnable Tree Filter for Generic Feature Transform](#)

- Lin Song, Yanwei Li, Zhengkai Jiang, Zeming Li, Xiangyu Zhang, Hongbin Sun, Jian Sun, Nanning Zheng
- abstract@[open-review](#): The Learnable Tree Filter presents a remarkable approach to model structure-preserving relations for semantic segmentation. Nevertheless, the intrinsic geometric constraint forces it to focus on the regions with close spatial distance, hindering the effective long-range interactions. To relax the geometric constraint, we give the analysis by reformulating it as a Markov Random Field and introduce a learnable unary term. Besides, we propose a learnable spanning tree algorithm to replace the original non-differentiable one, which further improves the flexibility and robustness. With the above improvements, our method can better capture long range dependencies and preserve structural details with linear complexity, which is extended to several vision tasks for more generic feature transform. Extensive experiments on object detection/instance segmentation demonstrate the consistent improvements over the original version. For semantic segmentation, we achieve leading performance (82.1% mIoU) on the Cityscapes benchmark without bells-and whistles. Code is available at <https://github.com/StevenGrove/LearnableTreeFilterV2>.

[Self-Supervised Relational Reasoning for Representation Learning](#)

- Massimiliano Patacchiola, Amos J. Storkey
- abstract@[open-review](#): In self-supervised learning, a system is tasked with achieving a surrogate objective by defining alternative targets on a set of unlabeled data. The aim is to build useful representations that can be used in downstream tasks, without costly manual annotation. In this work, we propose a novel self-supervised formulation of relational reasoning that allows a learner to bootstrap a signal from information implicit in unlabeled data. Training a relation head to discriminate how entities relate to themselves (intra-reasoning) and other entities (inter-reasoning), results in rich and descriptive representations in the underlying neural network backbone, which can be used in downstream tasks such as classification and image retrieval. We evaluate the proposed method following a rigorous experimental procedure, using standard datasets, protocols, and backbones. Self-supervised relational reasoning outperforms the best competitor in all conditions by an average 14% in accuracy, and the most recent state-of-the-art model by 3%. We link the effectiveness of the method to the maximization of a Bernoulli log-likelihood, which can be considered as a proxy for maximizing the mutual information, resulting in a more efficient objective with respect to the commonly used contrastive losses.

[Sufficient dimension reduction for classification using principal optimal transport direction](#)

- Cheng Meng, Jun Yu, Jingyi Zhang, Ping Ma, Wenxuan Zhong
- abstract@[open-review](#): Sufficient dimension reduction is used pervasively as a supervised dimension reduction approach. Most existing sufficient dimension reduction methods are developed for data with a continuous response and may have an unsatisfactory performance for the categorical response, especially for the binary-response. To address this issue, we propose a novel estimation method of sufficient dimension reduction subspace (SDR subspace) using optimal transport. The proposed method, named principal optimal transport direction (POTD), estimates the basis of the SDR subspace using the principal directions of the optimal transport coupling between the data respecting different response categories. The proposed method also reveals the relationship among three seemingly irrelevant topics, i.e., sufficient dimension reduction, support vector machine, and optimal transport. We study the asymptotic properties of POTD and show that in the cases when the class labels contain no error, POTD estimates the SDR subspace exclusively. Empirical studies show POTD outperforms most of the state-of-the-art linear dimension reduction methods.

[Fast Epigraphical Projection-based Incremental Algorithms for Wasserstein Distributionally Robust Support Vector Machine](#)

- Jiajin Li, Caihua Chen, Anthony Man-Cho So
- abstract@[open-review](#): Wasserstein distributionally robust optimization (DRO) is concerned with finding decisions that perform well on data that are drawn from the worst probability distribution within a Wasserstein ball centered at a certain nominal distribution. In recent years, it has been shown that various DRO formulations of learning models admit tractable convex reformulations. However, most existing works propose to solve these convex reformulations by general-purpose solvers, which are not well-suited for tackling large-scale problems. In this paper, we focus on a family of Wasserstein distributionally robust support vector machine (DRSVM) problems and propose two novel epigraphical projection-based incremental algorithms to solve them. The updates in each iteration of these algorithms can be computed in a highly efficient manner. Moreover, we show that the DRSVM problems considered in this paper satisfy a Hölderian growth condition with explicitly determined growth exponents. Consequently, we are able to establish the convergence rates of the proposed incremental algorithms. Our numerical results indicate that the proposed methods are orders of magnitude faster than the state-of-the-art, and the performance gap grows considerably as the problem size increases.

[Differentially Private Clustering: Tight Approximation Ratios](#)

- Badih Ghazi, Ravi Kumar, Pasin Manurangsi
- abstract@[open-review](#): We study the task of differentially private clustering. For several basic clustering problems, including Euclidean DensestBall, 1-Cluster, k-means, and k-median, we give efficient differentially private algorithms that achieve essentially the same approximation ratios as those that can be obtained by any non-private algorithm, while incurring only small additive errors. This improves upon existing efficient algorithms that only achieve some large constant approximation factors. Our results also imply an improved algorithm for the Sample and Aggregate privacy framework. Furthermore, we show that one of the tools used in our 1-Cluster algorithm can be employed to get a faster quantum algorithm for ClosestPair in a moderate number of dimensions.

[On the Power of Louvain in the Stochastic Block Model](#)

- Vincent Cohen-Addad, Adrian Kosowski, Frederik Mallmann-Trenn, David Saulpic
- abstract@[open-review](#): A classic problem in machine learning and data analysis is to partition the vertices of a network in such a way that vertices in the same set are densely connected and vertices in different sets are loosely connected. In practice, the most popular approaches rely on local search algorithms; not only for the ease of implementation and the efficiency, but also because of the accuracy of these methods on many real world graphs. For

example, the Louvain algorithm -- a local search based algorithm -- has quickly become the method of choice for clustering in social networks. However, explaining the success of these methods remains an open problem: in the worst-case, the runtime can be up to $\Omega(n^2)$, much worse than what is typically observed in practice, and no guarantee on the quality of its output can be established. The goal of this paper is to shed light on the inner-workings of Louvain; only if we understand Louvain, can we rely on it and further improve it. To achieve this goal, we study the behavior of Louvain in the famous two-bloc Stochastic Block Model, which has a clear ground-truth and serves as the standard testbed for graph clustering algorithms. We provide valuable tools for the analysis of Louvain, but also for many other combinatorial algorithms. For example, we show that the probability for a node to have more edges towards its own community is $1/2 + \Omega(\min(\Delta(p-q)\sqrt{np}, 1))$ in the SBM(n,p,q), where Δ is the imbalance. Note that this bound is asymptotically tight and useful for the analysis of a wide range of algorithms (Louvain, Kernighan-Lin, Simulated Annealing etc).

[Fairness with Overlapping Groups; a Probabilistic Perspective](#)

- Forest Yang, Mouhamadou Cisse, Sanmi Koyejo
- abstract@[open-review](#): In algorithmically fair prediction problems, a standard goal is to ensure the equality of fairness metrics across multiple overlapping groups simultaneously. We reconsider this standard fair classification problem using a probabilistic population analysis, which, in turn, reveals the Bayes-optimal classifier. Our approach unifies a variety of existing group-fair classification methods and enables extensions to a wide range of non-decomposable multiclass performance metrics and fairness measures. The Bayes-optimal classifier further inspires consistent procedures for algorithmically fair classification with overlapping groups. On a variety of real datasets, the proposed approach outperforms baselines in terms of its fairness-performance tradeoff.

[AttendLight: Universal Attention-Based Reinforcement Learning Model for Traffic Signal Control](#)

- Afshin Oroojlooy, Mohammadreza Nazari, Davood Hajinezhad, Jorge Silva
- abstract@[open-review](#): We propose AttendLight, an end-to-end Reinforcement Learning (RL) algorithm for the problem of traffic signal control. Previous approaches for this problem have the shortcoming that they require training for each new intersection with a different structure or traffic flow distribution. AttendLight solves this issue by training a single, universal model for intersections with any number of roads, lanes, phases (possible signals), and traffic flow. To this end, we propose a deep RL model which incorporates two attention models. The first attention model is introduced to handle different numbers of roads-lanes; and the second attention model is intended for enabling decision-making with any number of phases in an intersection. As a result, our proposed model works for any intersection configuration, as long as a similar configuration is represented in the training set. Experiments were conducted with both synthetic and real-world standard benchmark datasets. Our numerical experiment covers intersections with three or four approaching roads; one-directional/bi-directional roads with one, two, and three lanes; different number of phases; and different traffic flows. We consider two regimes: (i) single-environment training, single-deployment, and (ii) multi-environment training, multi-deployment. AttendLight outperforms both classical and other RL-based approaches on all cases in both regimes.

[Searching for Low-Bit Weights in Quantized Neural Networks](#)

- Zhaohui Yang, Yunhe Wang, Kai Han, Chunjing XU, Chao Xu, Dacheng Tao, Chang Xu
- abstract@[open-review](#): Quantized neural networks with low-bit weights and activations are attractive for developing AI accelerators. However, the quantization functions used in most conventional quantization methods are non-differentiable, which increases the optimization difficulty of quantized networks. Compared with full-precision parameters (i.e., 32-bit floating numbers), low-bit values are selected from a much smaller set. For example, there are only 16 possibilities in 4-bit space. Thus, we present to regard the discrete weights in an arbitrary quantized neural network as searchable variables, and utilize a differential method to search them accurately. In particular, each weight is represented as a probability distribution over the discrete value set. The probabilities are optimized during training and the values with the highest probability are selected to establish the desired quantized network. Experimental results on benchmarks demonstrate that the proposed method is able to produce quantized neural networks with higher performance over the state-of-the-arts on both image classification and super-resolution tasks.

[Adaptive Reduced Rank Regression](#)

- Qiong Wu, Felix MF Wong, Yanhua Li, Zhenming Liu, Varun Kanade
- abstract@[open-review](#): We study the low rank regression problem $y = Mx + \hat{\mu}$, where x and y are d_1 and d_2 dimensional vectors respectively. We consider the extreme high-dimensional setting where the number of observations n is less than $d_1 + d_2$. Existing algorithms are designed for settings where n is typically as large as $\text{rank}(M)(d_1+d_2)$. This work provides an efficient algorithm which only involves two SVD, and establishes statistical guarantees on its performance. The algorithm decouples the problem by first estimating the precision matrix of the features, and then solving the matrix denoising problem. To complement the upper bound, we introduce new techniques for establishing lower bounds on the performance of any algorithm for this problem. Our preliminary experiments confirm that our algorithm often out-performs existing baseline, and is always at least competitive.

[From Predictions to Decisions: Using Lookahead Regularization](#)

- Nir Rosenfeld, Anna Hilgard, Sai Srivatsa Ravindranath, David C. Parkes
- abstract@[open-review](#): Machine learning is a powerful tool for predicting human-related outcomes, from creditworthiness to heart attack risks. But when deployed transparently, learned models also affect how users act in order to improve outcomes. The standard approach to learning predictive models is agnostic to induced user actions and provides no guarantees as to the effect of actions. We provide a framework for learning predictors that are accurate, while also considering interactions between the learned model and user decisions. For this, we introduce look-ahead regularization which, by anticipating user actions, encourages predictive models to also induce actions that improve outcomes. This regularization carefully tailors the uncertainty estimates that govern confidence in this improvement to the distribution of model-induced actions. We report the results of experiments on real and synthetic data that show the effectiveness of this approach.

[Sequential Bayesian Experimental Design with Variable Cost Structure](#)

- Sue Zheng, David Hayden, Jason Pacheco, John W. Fisher III
- abstract@[open-review](#): Mutual information (MI) is a commonly adopted utility function in Bayesian optimal experimental design (BOED). While theoretically appealing, MI evaluation poses a significant computational burden for most real world applications. As a result, many algorithms utilize MI bounds as proxies that lack regret-style guarantees. Here, we utilize two-sided bounds to provide such guarantees. Bounds are successively refined/tightened through additional computation until a desired guarantee is achieved. We consider the problem of adaptively allocating computational resources in BOED. Our approach achieves the same guarantee as existing methods, but with fewer evaluations of the costly MI reward. We adapt knapsack optimization of best arm identification problems, with important differences that impact overall algorithm design and performance. First, observations of MI rewards are biased. Second, evaluating experiments incurs shared costs amongst all experiments (posterior sampling) in addition to per experiment costs that may vary with increasing evaluation. We propose and demonstrate an algorithm that accounts for these variable costs in the refinement decision.

[Predictive inference is free with the jackknife+-after-bootstrap](#)

- Byol Kim, Chen Xu, Rina Barber
- abstract@[open-review](#): Ensemble learning is widely used in applications to make predictions in complex decision problems---for example, averaging models fitted to a sequence of samples bootstrapped from the available training data. While such methods offer more accurate, stable, and robust predictions and model estimates, much less is known about how to perform valid, assumption-lean inference on the output of these types of procedures. In this paper, we propose the jackknife+-after-bootstrap (J+aB), a procedure for constructing a predictive interval, which uses only the available bootstrapped samples and their corresponding fitted models, and is therefore "free" in terms of the cost of model fitting. The J+aB offers a predictive coverage guarantee that holds with no assumptions on the distribution of the data, the nature of the fitted model, or the way in which the ensemble of models are aggregated---at worst, the failure rate of the predictive interval is inflated by a factor of 2. Our numerical experiments verify the coverage and accuracy of the resulting predictive intervals on real data.

[Counterfactual Predictions under Runtime Confounding](#)

- Amanda Coston, Edward Kennedy, Alexandra Chouldechova
- abstract@[open-review](#): Algorithms are commonly used to predict outcomes under a particular decision or intervention, such as predicting likelihood of default if a loan is approved. Generally, to learn such counterfactual prediction models from observational data on historical decisions and corresponding outcomes, one must measure all factors that jointly affect the outcome and the decision taken. Motivated by decision support applications, we study the counterfactual prediction task in the setting where all relevant factors are captured in the historical data, but it is infeasible, undesirable, or impermissible to use some such factors in the prediction model. We refer to this setting as runtime confounding. We propose a doubly-robust procedure for learning counterfactual prediction models in this setting. Our theoretical analysis and experimental results suggest that our method often outperforms competing approaches. We also present a validation procedure for evaluating the performance of counterfactual prediction methods.

[Learning Loss for Test-Time Augmentation](#)

- Ildoo Kim, Younghoon Kim, Sungwoong Kim
- abstract@[open-review](#): Data augmentation has been actively studied for robust neural networks. Most of the recent data augmentation methods focus on augmenting datasets during the training phase. At the testing phase, simple transformations are still widely used for test-time augmentation. This paper proposes a novel instance-level test- time augmentation that efficiently selects suitable transformations for a test input. Our proposed method involves an auxiliary module to predict the loss of each possible transformation given the input. Then, the transformations having lower predicted losses are applied to the input. The network obtains the results by averaging the prediction results of augmented inputs. Experimental results on several image classification benchmarks show that the proposed instance-aware test- time augmentation improves the model's robustness against various corruptions.

[Balanced Meta-Softmax for Long-Tailed Visual Recognition](#)

- Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, hongsheng Li
- abstract@[open-review](#): Deep classifiers have achieved great success in visual recognition. However, real-world data is long-tailed by nature, leading to the mismatch between training and testing distributions. In this paper, we show that the Softmax function, though used in most classification tasks, gives a biased gradient estimation under the long-tailed setup. This paper presents Balanced Softmax, an elegant unbiased extension of Softmax, to accommodate the label distribution shift between training and testing. Theoretically, we derive the generalization bound for multiclass Softmax regression and show our loss minimizes the bound. In addition, we introduce Balanced Meta-Softmax, applying a complementary Meta Sampler to estimate the optimal class sample rate and further improve long-tailed learning. In our experiments, we demonstrate that Balanced Meta-Softmax outperforms state-of-the-art long-tailed classification solutions on both visual recognition and instance segmentation tasks.

[Efficient Exploration of Reward Functions in Inverse Reinforcement Learning via Bayesian Optimization](#)

- Sreejith Balakrishnan, Quoc Phong Nguyen, Bryan Kian Hsiang Low, Harold Soh
- abstract@[open-review](#): The problem of inverse reinforcement learning (IRL) is relevant to a variety of tasks including value alignment and robot learning from demonstration. Despite significant algorithmic contributions in recent years, IRL remains an ill-posed problem at its core; multiple reward functions coincide with the observed behavior and the actual reward function is not identifiable without prior knowledge or supplementary information. This paper presents an IRL framework called Bayesian optimization-IRL (BO-IRL) which identifies multiple solutions that are consistent with the expert demonstrations by efficiently exploring the reward function space. BO-IRL achieves this by utilizing Bayesian Optimization along with our newly proposed kernel that (a) projects the parameters of policy invariant reward functions to a single point in a latent space and (b) ensures nearby points in the latent space correspond to reward functions yielding similar likelihoods. This projection allows the use of standard stationary kernels in the latent space to capture the correlations present across the reward function space. Empirical results on synthetic and real-world environments (model-free and model-based) show that BO-IRL discovers multiple reward functions while minimizing the number of expensive exact policy optimizations.

[MDP Homomorphic Networks: Group Symmetries in Reinforcement Learning](#)

- Elise van der Pol, Daniel Worrall, Herke van Hoof, Frans Oliehoek, Max Welling
- abstract@[open-review](#): This paper introduces MDP homomorphic networks for deep reinforcement learning. MDP homomorphic networks are neural networks that are equivariant under symmetries in the joint state-action space of an MDP. Current approaches to deep reinforcement learning do not usually exploit knowledge about such structure. By building this prior knowledge into policy and value networks using an equivariance constraint, we can reduce the size of the solution space. We specifically focus on group-structured symmetries (invertible transformations). Additionally, we introduce an easy method for constructing equivariant network layers numerically, so the system designer need not solve the constraints by hand, as is typically done. We construct MDP homomorphic MLPs and CNNs that are equivariant under either a group of reflections or rotations. We show that such networks converge faster than unstructured baselines on CartPole, a grid world and Pong.

[How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods](#)

- Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, Mani Srivastava
- abstract@[open-review](#): Explaining the inner workings of deep neural network models have received considerable attention in recent years. Researchers have attempted to provide human parseable explanations justifying why a model performed a specific classification. Although many of these toolkits are available for use, it is unclear which style of explanation is preferred by end-users, thereby demanding investigation. We performed a cross-analysis Amazon Mechanical Turk study comparing the popular state-of-the-art explanation methods to empirically determine which are better in explaining model decisions. The participants were asked to compare explanation methods across applications spanning image, text, audio, and sensory domains. Among the surveyed methods, explanation-by-example was preferred in all domains except text sentiment classification, where LIME's method of annotating input text was preferred. We highlight qualitative aspects of employing the studied explainability methods and conclude with implications for researchers and engineers that seek to incorporate explanations into user-facing deployments.

[On the Error Resistance of Hinge-Loss Minimization](#)

- Kunal Talwar

- abstract@[open-review](#): Commonly used classification algorithms in machine learning, such as support vector machines, minimize a convex surrogate loss on training examples. In practice, these algorithms are surprisingly robust to errors in the training data. In this work, we identify a set of conditions on the data under which such surrogate loss minimization algorithms provably learn the correct classifier. This allows us to establish, in a unified framework, the robustness of these algorithms under various models on data as well as error. In particular, we show that if the data is linearly classifiable with a slightly non-trivial margin (i.e. a margin at least $C \cdot \text{div}(\sqrt{d})$ for d -dimensional unit vectors), and the class-conditional distributions are near isotropic and logconcave, then surrogate loss minimization has negligible error on the uncorrupted data even when a constant fraction of examples are adversarially mislabeled.

[Munchausen Reinforcement Learning](#)

- Nino Vieillard, Olivier Pietquin, Matthieu Geist
- abstract@[open-review](#): Bootstrapping is a core mechanism in Reinforcement Learning (RL). Most algorithms, based on temporal differences, replace the true value of a transiting state by their current estimate of this value. Yet, another estimate could be leveraged to bootstrap RL: the current policy. Our core contribution stands in a very simple idea: adding the scaled log-policy to the immediate reward. We show that, by slightly modifying Deep Q-Network (DQN) in that way provides an agent that is competitive with the state-of-the-art Rainbow on Atari games, without making use of distributional RL, n-step returns or prioritized replay. To demonstrate the versatility of this idea, we also use it together with an Implicit Quantile Network (IQN). The resulting agent outperforms Rainbow on Atari, installing a new State of the Art with very little modifications to the original algorithm. To add to this empirical study, we provide strong theoretical insights on what happens under the hood -- implicit Kullback-Leibler regularization and increase of the action-gap.

[Object Goal Navigation using Goal-Oriented Semantic Exploration](#)

- Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, Russ R. Salakhutdinov
- abstract@[open-review](#): This work studies the problem of object goal navigation which involves navigating to an instance of the given object category in unseen environments. End-to-end learning-based navigation methods struggle at this task as they are ineffective at exploration and long-term planning. We propose a modular system called, 'Goal-Oriented Semantic Exploration' which builds an episodic semantic map and uses it to explore the environment efficiently based on the goal object category. Empirical results in visually realistic simulation environments show that the proposed model outperforms a wide range of baselines including end-to-end learning-based methods as well as modular map-based methods and led to the winning entry of the CVPR-2020 Habitat ObjectNav Challenge. Ablation analysis indicates that the proposed model learns semantic priors of the relative arrangement of objects in a scene, and uses them to explore efficiently. Domain-agnostic module design allows us to transfer our model to a mobile robot platform and achieve similar performance for object goal navigation in the real-world.

[Efficient semidefinite-programming-based inference for binary and multi-class MRFs](#)

- Chirag Pabbaraju, Po-Wei Wang, J. Zico Kolter
- abstract@[open-review](#): Probabilistic inference in pairwise Markov Random Fields (MRFs), i.e. computing the partition function or computing a MAP estimate of the variables, is a foundational problem in probabilistic graphical models. Semidefinite programming relaxations have long been a theoretically powerful tool for analyzing properties of probabilistic inference, but have not been practical owing to the high computational cost of typical solvers for solving the resulting SDPs. In this paper, we propose an efficient method for computing the partition function or MAP estimate in a pairwise MRF by instead exploiting a recently proposed coordinate-descent-based fast semidefinite solver. We also extend semidefinite relaxations from the typical binary MRF to the full multi-class setting, and develop a compact semidefinite relaxation that can again be solved efficiently using the solver. We show that the method substantially outperforms (both in terms of solution quality and speed) the existing state of the art in approximate inference, on benchmark problems drawn from previous work. We also show that our approach can scale to large MRF domains such as fully-connected pairwise CRF models used in computer vision.

[Funnel-Transformer: Filtering out Sequential Redundancy for Efficient Language Processing](#)

- Zihang Dai, Guokun Lai, Yiming Yang, Quoc Le
- abstract@[open-review](#): With the success of language pretraining, it is highly desirable to develop more efficient architectures of good scalability that can exploit the abundant unlabeled data at a lower cost. To improve the efficiency, we examine the much-overlooked redundancy in maintaining a full-length token-level presentation, especially for tasks that only require a single-vector presentation of the sequence. With this intuition, we propose Funnel-Transformer which gradually compresses the sequence of hidden states to a shorter one and hence reduces the computation cost. More importantly, by re-investing the saved FLOPs from length reduction in constructing a deeper or wider model, we further improve the model capacity. In addition, to perform token-level predictions as required by common pretraining objectives, Funnel-Transformer is able to recover a deep representation for each token from the reduced hidden sequence via a decoder. Empirically, with comparable or fewer FLOPs, Funnel-Transformer outperforms the standard Transformer on a wide variety of sequence-level prediction tasks, including text classification, language understanding, and reading comprehension.

[Semantic Visual Navigation by Watching YouTube Videos](#)

- Matthew Chang, Arjun Gupta, Saurabh Gupta
- abstract@[open-review](#): Semantic cues and statistical regularities in real-world environment layouts can improve efficiency for navigation in novel environments. This paper learns and leverages such semantic cues for navigating to objects of interest in novel environments, by simply watching YouTube videos. This is challenging because YouTube videos don't come with labels for actions or goals, and may not even showcase optimal behavior. Our method tackles these challenges through the use of Q-learning on pseudo-labeled transition quadruples (image, action, next image, reward). We show that such off-policy Q-learning from passive data is able to learn meaningful semantic cues for navigation. These cues, when used in a hierarchical navigation policy, lead to improved efficiency at the ObjectGoal task in visually realistic simulations. We observe a relative improvement of 15-83% over end-to-end RL, behavior cloning, and classical methods, while using minimal direct interaction.

[Heavy-tailed Representations, Text Polarity Classification & Data Augmentation](#)

- Hamid Jalalzai, Pierre Colombo, Chloé Clavel, Eric Gaussier, Giovanna Varni, Emmanuel Vignon, Anne Sabourin
- abstract@[open-review](#): The dominant approaches to text representation in natural language rely on learning embeddings on massive corpora which have convenient properties such as compositionality and distance preservation. In this paper, we develop a novel method to learn a heavy-tailed embedding with desirable regularity properties regarding the distributional tails, which allows to analyze the points far away from the distribution bulk using the framework of multivariate extreme value theory. In particular, a classifier dedicated to the tails of the proposed embedding is obtained which exhibits a scale invariance property exploited in a novel text generation method for label preserving dataset augmentation. Experiments on synthetic and real text data show the relevance of the proposed framework and confirm that this method generates meaningful sentences with controllable attribute, e.g. positive or negative sentiments.

[SuperLoss: A Generic Loss for Robust Curriculum Learning](#)

- Thibault Castells, Philippe Weinzaepfel, Jerome Revaud

- abstract@[open-review](#): Curriculum learning is a technique to improve a model performance and generalization based on the idea that easy samples should be presented before difficult ones during training. While it is generally complex to estimate a priori the difficulty of a given sample, recent works have shown that curriculum learning can be formulated dynamically in a self-supervised manner. The key idea is to somehow estimate the importance (or weight) of each sample directly during training based on the observation that easy and hard samples behave differently and can therefore be separated. However, these approaches are usually limited to a specific task (e.g., classification) and require extra data annotations, layers or parameters as well as a dedicated training procedure. We propose instead a simple and generic method that can be applied to a variety of losses and tasks without any change in the learning procedure. It consists in appending a novel loss function on top of any existing task loss, hence its name: the SuperLoss. Its main effect is to automatically downweight the contribution of samples with a large loss, i.e. hard samples, effectively mimicking the core principle of curriculum learning. As a side effect, we show that our loss prevents the memorization of noisy samples, making it possible to train from noisy data even with non-robust loss functions. Experimental results on image classification, regression, object detection and image retrieval demonstrate consistent gain, particularly in the presence of noise.

[CogMol: Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models](#)

- Vijil Chenthamarakshan, Payel Das, Samuel Hoffman, Hendrik Strobelt, Inkit Padhi, Kar Wai Lim, Benjamin Hoover, Matteo Manica, Jannis Born, Teodoro Laino, Aleksandra Mojsilovic
- abstract@[open-review](#): The novel nature of SARS-CoV-2 calls for the development of efficient de novo drug design approaches. In this study, we propose an end-to-end framework, named CogMol (Controlled Generation of Molecules), for designing new drug-like small molecules targeting novel viral proteins with high affinity and off-target selectivity. CogMol combines adaptive pre-training of a molecular SMILES Variational Autoencoder (VAE) and an efficient multi-attribute controlled sampling scheme that uses guidance from attribute predictors trained on latent features. To generate novel and optimal drug-like molecules for unseen viral targets, CogMol leverages a protein-molecule binding affinity predictor that is trained using SMILES VAE embeddings and protein sequence embeddings learned unsupervised from a large corpus. We applied the CogMol framework to three SARS-CoV-2 target proteins: main protease, receptor-binding domain of the spike protein, and non-structural protein 9 replicase. The generated candidates are novel at both the molecular and chemical scaffold levels when compared to the training data. CogMol also includes insilico screening for assessing toxicity of parent molecules and their metabolites with a multi-task toxicity classifier, synthetic feasibility with a chemical retrosynthesis predictor, and target structure binding with docking simulations. Docking reveals favorable binding of generated molecules to the target protein structure, where 87--95% of high affinity molecules showed docking free energy $\$ < -6 \text{ kcal/mol}$. When compared to approved drugs, the majority of designed compounds show low predicted parent molecule and metabolite toxicity and high predicted synthetic feasibility. In summary, CogMol can handle multi-constraint design of synthesizable, low-toxic, drug-like molecules with high target specificity and selectivity, even to novel protein target sequences, and does not need target-dependent fine-tuning of the framework or target structure information.

[Memory Based Trajectory-conditioned Policies for Learning from Sparse Rewards](#)

- Yijie Guo, Jongwook Choi, Marcin Moczulski, Shengyu Feng, Samy Bengio, Mohammad Norouzi, Honglak Lee
- abstract@[open-review](#): Reinforcement learning with sparse rewards is challenging because an agent can rarely obtain non-zero rewards and hence, gradient-based optimization of parameterized policies can be incremental and slow. Recent work demonstrated that using a memory buffer of previous successful trajectories can result in more effective policies. However, existing methods may overly exploit past successful experiences, which can encourage the agent to adopt sub-optimal and myopic behaviors. In this work, instead of focusing on good experiences with limited diversity, we propose to learn a trajectory-conditioned policy to follow and expand diverse past trajectories from a memory buffer. Our method allows the agent to reach diverse regions in the state space and improve upon the past trajectories to reach new states. We empirically show that our approach significantly outperforms count-based exploration methods (parametric approach) and self-imitation learning (parametric approach with non-parametric memory) on various complex tasks with local optima. In particular, without using expert demonstrations or resetting to arbitrary states, we achieve the state-of-the-art scores under five billion number of frames, on challenging Atari games such as Montezuma's Revenge and Pitfall.

[Liberty or Depth: Deep Bayesian Neural Nets Do Not Need Complex Weight Posterior Approximations](#)

- Sebastian Farquhar, Lewis Smith, Yarin Gal
- abstract@[open-review](#): We challenge the longstanding assumption that the mean-field approximation for variational inference in Bayesian neural networks is severely restrictive, and show this is not the case in deep networks. We prove several results indicating that deep mean-field variational weight posteriors can induce similar distributions in function-space to those induced by shallower networks with complex weight posteriors. We validate our theoretical contributions empirically, both through examination of the weight posterior using Hamiltonian Monte Carlo in small models and by comparing diagonal- to structured-covariance in large settings. Since complex variational posteriors are often expensive and cumbersome to implement, our results suggest that using mean-field variational inference in a deeper model is both a practical and theoretically justified alternative to structured approximations.

[Improving Sample Complexity Bounds for \(Natural\) Actor-Critic Algorithms](#)

- Tengyu Xu, Zhe Wang, Yingbin Liang
- abstract@[open-review](#): The actor-critic (AC) algorithm is a popular method to find an optimal policy in reinforcement learning. In the infinite horizon scenario, the finite-sample convergence rate for the AC and natural actor-critic (NAC) algorithms has been established recently, but under independent and identically distributed (i.i.d.) sampling and single-sample update at each iteration. In contrast, this paper characterizes the convergence rate and sample complexity of AC and NAC under Markovian sampling, with mini-batch data for each iteration, and with actor having general policy class approximation. We show that the overall sample complexity for a mini-batch AC to attain an $\$ \epsilon \$$ -accurate stationary point improves the best known sample complexity of AC by an order of $\$ \mathcal{O}(\epsilon^{-1} \log(1/\epsilon)) \$$, and the overall sample complexity for a mini-batch NAC to attain an $\$ \epsilon \$$ -accurate globally optimal point improves the existing sample complexity of NAC by an order of $\$ \mathcal{O}(\epsilon^{-2} \log(1/\epsilon)) \$$. Moreover, the sample complexity of AC and NAC characterized in this work outperforms that of policy gradient (PG) and natural policy gradient (NPG) by a factor of $\$ \mathcal{O}((1-\gamma)^{-3}) \$$ and $\$ \mathcal{O}((1-\gamma)^{-4} \epsilon^{-2} \log(1/\epsilon)) \$$, respectively. This is the first theoretical study establishing that AC and NAC attain otherwise performance improvement over PG and NPG under infinite horizon due to the incorporation of critic.

[Learning Differential Equations that are Easy to Solve](#)

- Jacob Kelly, Jesse Bettencourt, Matthew J. Johnson, David K. Duvenaud
- abstract@[open-review](#): Differential equations parameterized by neural networks become expensive to solve numerically as training progresses. We propose a remedy that encourages learned dynamics to be easier to solve. Specifically, we introduce a differentiable surrogate for the time cost of standard numerical solvers, using higher-order derivatives of solution trajectories. These derivatives are efficient to compute with Taylor-mode automatic differentiation. Optimizing this additional objective trades model performance against the time cost of solving the learned dynamics. We demonstrate our approach by training substantially faster, while nearly as accurate, models in supervised classification, density estimation, and time-series modelling tasks.

[Stability of Stochastic Gradient Descent on Nonsmooth Convex Losses](#)

- Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, Kunal Talwar

- abstract@[open-review](#): Uniform stability is a notion of algorithmic stability that bounds the worst case change in the model output by the algorithm when a single data point in the dataset is replaced. An influential work of Hardt et al. [2016] provides strong upper bounds on the uniform stability of the stochastic gradient descent (SGD) algorithm on sufficiently smooth convex losses. These results led to important progress in understanding of the generalization properties of SGD and several applications to differentially private convex optimization for smooth losses. Our work is the first to address uniform stability of SGD on nonsmooth convex losses. Specifically, we provide sharp upper and lower bounds for several forms of SGD and full-batch GD on arbitrary Lipschitz nonsmooth convex losses. Our lower bounds show that, in the nonsmooth case, (S)GD can be inherently less stable than in the smooth case. On the other hand, our upper bounds show that (S)GD is sufficiently stable for deriving new and useful bounds on generalization error. Most notably, we obtain the first dimension-independent generalization bounds for multi-pass SGD in the nonsmooth case. In addition, our bound allow us to derive a new algorithm for differentially private nonsmooth stochastic convex optimization with optimal excess population risk. Our algorithm is simpler and more efficient than the best known algorithm for the nonsmooth case, due to Feldman et al. [2020].

[Influence-Augmented Online Planning for Complex Environments](#)

- Jinke He, Miguel Suau de Castro, Frans Oliehoek
- abstract@[open-review](#): How can we plan efficiently in real time to control an agent in a complex environment that may involve many other agents? While existing sample-based planners have enjoyed empirical success in large POMDPs, their performance heavily relies on a fast simulator. However, real-world scenarios are complex in nature and their simulators are often computationally demanding, which severely limits the performance of online planners. In this work, we propose influence-augmented online planning, a principled method to transform a factored simulator of the entire environment into a local simulator that samples only the state variables that are most relevant to the observation and reward of the planning agent and captures the incoming influence from the rest of the environment using machine learning methods. Our main experimental results show that planning on this less accurate but much faster local simulator with POMCP leads to higher real-time planning performance than planning on the simulator that models the entire environment.

[PAC-Bayes Learning Bounds for Sample-Dependent Priors](#)

- Pranjal Awasthi, Satyen Kale, Stefani Karp, Mehryar Mohri
- abstract@[open-review](#): We present a series of new PAC-Bayes learning guarantees for randomized algorithms with sample-dependent priors. Our most general bounds make no assumption on the priors and are given in terms of certain covering numbers under the infinite-Renyi divergence and the L1 distance. We show how to use these general bounds to derive leaning bounds in the setting where the sample-dependent priors obey an infinite-Renyi divergence or L1-distance sensitivity condition. We also provide a flexible framework for computing PAC-Bayes bounds, under certain stability assumptions on the sample-dependent priors, and show how to use this framework to give more refined bounds when the priors satisfy an infinite-Renyi divergence sensitivity condition.

[Reward-rational \(implicit\) choice: A unifying formalism for reward learning](#)

- Hong Jun Jeon, Smitha Milli, Anca Dragan
- abstract@[open-review](#): It is often difficult to hand-specify what the correct reward function is for a task, so researchers have instead aimed to learn reward functions from human behavior or feedback. The types of behavior interpreted as evidence of the reward function have expanded greatly in recent years. We've gone from demonstrations, to comparisons, to reading into the information leaked when the human is pushing the robot away or turning it off. And surely, there is more to come. How will a robot make sense of all these diverse types of behavior? Our key observation is that different types of behavior can be interpreted in a single unifying formalism - as a reward-rational choice that the human is making, often implicitly. We use this formalism to survey prior work through a unifying lens, and discuss its potential use as a recipe for interpreting new sources of information that are yet to be uncovered.

[Probabilistic Time Series Forecasting with Shape and Temporal Diversity](#)

- Vincent LE GUEN, Nicolas THOME
- abstract@[open-review](#): Probabilistic forecasting consists in predicting a distribution of possible future outcomes. In this paper, we address this problem for non-stationary time series, which is very challenging yet crucially important. We introduce the STRIPE model for representing structured diversity based on shape and time features, ensuring both probable predictions while being sharp and accurate. STRIPE is agnostic to the forecasting model, and we equip it with a diversification mechanism relying on determinantal point processes (DPP). We introduce two DPP kernels for modelling diverse trajectories in terms of shape and time, which are both differentiable and proved to be positive semi-definite. To have an explicit control on the diversity structure, we also design an iterative sampling mechanism to disentangle shape and time representations in the latent space. Experiments carried out on synthetic datasets show that STRIPE significantly outperforms baseline methods for representing diversity, while maintaining accuracy of the forecasting model. We also highlight the relevance of the iterative sampling scheme and the importance to use different criteria for measuring quality and diversity. Finally, experiments on real datasets illustrate that STRIPE is able to outperform state-of-the-art probabilistic forecasting approaches in the best sample prediction.

[Low Distortion Block-Resampling with Spatially Stochastic Networks](#)

- Sarah Hong, Martin Arjovsky, Darryl Barnhart, Ian Thompson
- abstract@[open-review](#): We formalize and attack the problem of generating new images from old ones that are as diverse as possible, only allowing them to change without restrictions in certain parts of the image while remaining globally consistent. This encompasses the typical situation found in generative modelling, where we are happy with parts of the generated data, but would like to resample others ("I like this generated castle overall, but this tower looks unrealistic, I would like a new one"). In order to attack this problem we build from the best conditional and unconditional generative models to introduce a new network architecture, training procedure, and a new algorithm for resampling parts of the image as desired.

[Continual Deep Learning by Functional Regularisation of Memorable Past](#)

- Pingbo Pan, Siddharth Swaroop, Alexander Immer, Runa Eschenhagen, Richard Turner, Mohammad Emtiyaz E. Khan
- abstract@[open-review](#): Continually learning new skills is important for intelligent systems, yet standard deep learning methods suffer from catastrophic forgetting of the past. Recent works address this with weight regularisation. Functional regularisation, although computationally expensive, is expected to perform better, but rarely does so in practice. In this paper, we fix this issue by using a new functional-regularisation approach that utilises a few memorable past examples crucial to avoid forgetting. By using a Gaussian Process formulation of deep networks, our approach enables training in weight-space while identifying both the memorable past and a functional prior. Our method achieves state-of-the-art performance on standard benchmarks and opens a new direction for life-long learning where regularisation and memory-based methods are naturally combined.

[Distance Encoding: Design Provably More Powerful Neural Networks for Graph Representation Learning](#)

- Pan Li, Yanbang Wang, Hongwei Wang, Jure Leskovec
- abstract@[open-review](#): Learning representations of sets of nodes in a graph is crucial for applications ranging from node-role discovery to link prediction and molecule classification. Graph Neural Networks (GNNs) have achieved great success in graph representation learning. However, expressive power of

GNNs is limited by the 1-Weisfeiler-Lehman (WL) test and thus GNNs generate identical representations for graph substructures that may in fact be very different. More powerful GNNs, proposed recently by mimicking higher-order-WL tests, only focus on representing entire graphs and they are computationally inefficient as they cannot utilize sparsity of the underlying graph. Here we propose and mathematically analyze a general class of structure-related features, termed Distance Encoding (DE). DE assists GNNs in representing any set of nodes, while providing strictly more expressive power than the 1-WL test. DE captures the distance between the node set whose representation is to be learned and each node in the graph. To capture the distance DE can apply various graph-distance measures such as shortest path distance or generalized PageRank scores. We propose two ways for GNNs to use DEs (1) as extra node features, and (2) as controllers of message aggregation in GNNs. Both approaches can utilize the sparse structure of the underlying graph, which leads to computational efficiency and scalability. We also prove that DE can distinguish node sets embedded in almost all regular graphs where traditional GNNs always fail. We evaluate DE on three tasks over six real networks: structural role prediction, link prediction, and triangle prediction. Results show that our models outperform GNNs without DE by up-to 15% in accuracy and AUROC. Furthermore, our models also significantly outperform other state-of-the-art methods especially designed for the above tasks.

[Fast Fourier Convolution](#)

- Lu Chi, Borui Jiang, Yadong Mu
- abstract@[open-review](#): Vanilla convolutions in modern deep networks are known to operate locally and at fixed scale (e.g., the widely-adopted 3*3 kernels in image-oriented tasks). This causes low efficacy in connecting two distant locations in the network. In this work, we propose a novel convolutional operator dubbed as fast Fourier convolution (FFC), which has the main hallmarks of non-local receptive fields and cross-scale fusion within the convolutional unit. According to spectral convolution theorem in Fourier theory, point-wise update in the spectral domain globally affects all input features involved in Fourier transform, which sheds light on neural architectural design with non-local receptive field. Our proposed FFC is inspired to encapsulate three different kinds of computations in a single operation unit: a local branch that conducts ordinary small-kernel convolution, a semi-global branch that processes spectrally stacked image patches, and a global branch that manipulates image-level spectrum. All branches complementarily address different scales. A multi-branch aggregation step is included in FFC for cross-scale fusion. FFC is a generic operator that can directly replace vanilla convolutions in a large body of existing networks, without any adjustments and with comparable complexity metrics (e.g., FLOPs). We experimentally evaluate FFC in three major vision benchmarks (ImageNet for image recognition, Kinetics for video action recognition, MSCOCO for human keypoint detection). It consistently elevates accuracies in all above tasks by significant margins.

[Unsupervised Learning of Dense Visual Representations](#)

- Pedro O. O. Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, Aaron C. Courville
- abstract@[open-review](#): Contrastive self-supervised learning has emerged as a promising approach to unsupervised visual representation learning. In general, these methods learn global (image-level) representations that are invariant to different views (i.e., compositions of data augmentation) of the same image. However, many visual understanding tasks require dense (pixel-level) representations. In this paper, we propose View-Agnostic Dense Representation (VADeR) for unsupervised learning of dense representations. VADeR learns pixelwise representations by forcing local features to remain constant over different viewing conditions. Specifically, this is achieved through pixel-level contrastive learning: matching features (that is, features that describes the same location of the scene on different views) should be close in an embedding space, while non-matching features should be apart. VADeR provides a natural representation for dense prediction tasks and transfers well to downstream tasks. Our method outperforms ImageNet supervised pretraining (and strong unsupervised baselines) in multiple dense prediction tasks.

[Higher-Order Certification For Randomized Smoothing](#)

- Jeet Mohapatra, Ching-Yun Ko, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, Luca Daniel
- abstract@[open-review](#): Randomized smoothing is a recently proposed defense against adversarial attacks that has achieved state-of-the-art provable robustness against $\|\cdot\|_2$ perturbations. A number of works have extended the guarantees to other metrics, such as $\|\cdot\|_1$ or $\|\cdot\|_\infty$, by using different smoothing measures. Although the current framework has been shown to yield near-optimal $\|\cdot\|_p$ radii, the total safety region certified by the current framework can be arbitrarily small compared to the optimal. In this work, we propose a framework to improve the certified safety region for these smoothed classifiers without changing the underlying smoothing scheme. The theoretical contributions are as follows: 1) We generalize the certification for randomized smoothing by reformulating certified radius calculation as a nested optimization problem over a class of functions. 2) We provide a method to calculate the certified safety region using zeroth-order and first-order information for Gaussian-smoothed classifiers. We also provide a framework that generalizes the calculation for certification using higher-order information. 3) We design efficient, high-confidence estimators for the relevant statistics of the first-order information. Combining the theoretical contribution 2) and 3) allows us to certify safety region that are significantly larger than ones provided by the current methods. On CIFAR and Imagenet, the new regions achieve significant improvements on general $\|\cdot\|_1$ certified radii and on the $\|\cdot\|_2$ certified radii for color-space attacks ($\|\cdot\|_2$ perturbation restricted to only one color/channel) while also achieving smaller improvements on the general $\|\cdot\|_2$ certified radii. As discussed in the future works section, our framework can also provide a way to circumvent the current impossibility results on achieving higher magnitudes of certified radii without requiring the use of data-dependent smoothing techniques.

[Learning Structured Distributions From Untrusted Batches: Faster and Simpler](#)

- Sitan Chen, Jerry Li, Ankur Moitra
- abstract@[open-review](#): We revisit the problem of learning from untrusted batches introduced by Qiao and Valiant [QV17]. Recently, Jain and Orlitsky [JO19] gave a simple semidefinite programming approach based on the cut-norm that achieves essentially information-theoretically optimal error in polynomial time. Concurrently, Chen et al. [CLM19] considered a variant of the problem where $\frac{1}{4}$ is assumed to be structured, e.g. log-concave, monotone hazard rate, t-modal, etc. In this case, it is possible to achieve the same error with sample complexity sublinear in n , and they exhibited a quasi-polynomial time algorithm for doing so using Haar wavelets. In this paper, we find an appealing way to synthesize the techniques of [JO19] and [CLM19] to give the best of both worlds: an algorithm which runs in polynomial time and can exploit structure in the underlying distribution to achieve sublinear sample complexity. Along the way, we simplify the approach of [JO19] by avoiding the need for SDP rounding and giving a more direct interpretation of it through the lens of soft filtering, a powerful recent technique in high-dimensional robust estimation. We validate the usefulness of our algorithms in preliminary experimental evaluations.

[Hierarchical Quantized Autoencoders](#)

- Will Williams, Sam Ringer, Tom Ash, David MacLeod, Jamie Dougherty, John Hughes
- abstract@[open-review](#): Despite progress in training neural networks for lossy image compression, current approaches fail to maintain both perceptual quality and abstract features at very low bitrates. Encouraged by recent success in learning discrete representations with Vector Quantized Variational Autoencoders (VQ-VAEs), we motivate the use of a hierarchy of VQ-VAEs to attain high factors of compression. We show that the combination of stochastic quantization and hierarchical latent structure aids likelihood-based image compression. This leads us to introduce a novel objective for training hierarchical VQ-VAEs. Our resulting scheme produces a Markovian series of latent variables that reconstruct images of high-perceptual quality which retain semantically meaningful features. We provide qualitative and quantitative evaluations on the CelebA and MNIST datasets.

[Diversity can be Transferred: Output Diversification for White- and Black-box Attacks](#)

- Yusuke Tashiro, Yang Song, Stefano Ermon
- abstract@[open-review](#): Adversarial attacks often involve random perturbations of the inputs drawn from uniform or Gaussian distributions, e.g. to initialize optimization-based white-box attacks or generate update directions in black-box attacks. These simple perturbations, however, could be sub-optimal as they are agnostic to the model being attacked. To improve the efficiency of these attacks, we propose Output Diversified Sampling (ODS), a novel sampling strategy that attempts to maximize diversity in the target model's outputs among the generated samples. While ODS is a gradient-based strategy, the diversity offered by ODS is transferable and can be helpful for both white-box and black-box attacks via surrogate models. Empirically, we demonstrate that ODS significantly improves the performance of existing white-box and black-box attacks. In particular, ODS reduces the number of queries needed for state-of-the-art black-box attacks on ImageNet by a factor of two.

[POLY-HOOT: Monte-Carlo Planning in Continuous Space MDPs with Non-Asymptotic Analysis](#)

- Weichao Mao, Kaiqing Zhang, Qiaomin Xie, Tamer Basar
- abstract@[open-review](#): Monte-Carlo planning, as exemplified by Monte-Carlo Tree Search (MCTS), has demonstrated remarkable performance in applications with finite spaces. In this paper, we consider Monte-Carlo planning in an environment with continuous state-action spaces, a much less understood problem with important applications in control and robotics. We introduce POLY-HOOT, an algorithm that augments MCTS with a continuous armed bandit strategy named Hierarchical Optimistic Optimization (HOO) (Bubeck et al., 2011). Specifically, we enhance HOO by using an appropriate polynomial, rather than logarithmic, bonus term in the upper confidence bounds. Such a polynomial bonus is motivated by its empirical successes in AlphaGo Zero (Silver et al., 2017b), as well as its significant role in achieving theoretical guarantees of finite space MCTS (Shah et al., 2019). We investigate, for the first time, the regret of the enhanced HOO algorithm in non-stationary bandit problems. Using this result as a building block, we establish non-asymptotic convergence guarantees for POLY-HOOT: the value estimate converges to an arbitrarily small neighborhood of the optimal value function at a polynomial rate. We further provide experimental results that corroborate our theoretical findings.

[AvE: Assistance via Empowerment](#)

- Yuqing Du, Stas Tiomkin, Emre Kiciman, Daniel Polani, Pieter Abbeel, Anca Dragan
- abstract@[open-review](#): One difficulty in using artificial agents for human-assistive applications lies in the challenge of accurately assisting with a person's goal(s). Existing methods tend to rely on inferring the human's goal, which is challenging when there are many potential goals or when the set of candidate goals is difficult to identify. We propose a new paradigm for assistance by instead increasing the human's ability to control their environment, and formalize this approach by augmenting reinforcement learning with human empowerment. This task-agnostic objective increases the person's autonomy and ability to achieve any eventual state. We test our approach against assistance based on goal inference, highlighting scenarios where our method overcomes failure modes stemming from goal ambiguity or misspecification. As existing methods for estimating empowerment in continuous domains are computationally hard, precluding its use in real time learned assistance, we also propose an efficient empowerment-inspired proxy metric. Using this, we are able to successfully demonstrate our method in a shared autonomy user study for a challenging simulated teleoperation task with human-in-the-loop training.

[Variational Policy Gradient Method for Reinforcement Learning with General Utilities](#)

- Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, Mengdi Wang
- abstract@[open-review](#): In recent years, reinforcement learning systems with general goals beyond a cumulative sum of rewards have gained traction, such as in constrained problems, exploration, and acting upon prior experiences. In this paper, we consider policy optimization in Markov Decision Problems, where the objective is a general utility function of the state-action occupancy measure, which subsumes several of the aforementioned examples as special cases. Such generality invalidates the Bellman equation. As this means that dynamic programming no longer works, we focus on direct policy search. Analogously to the Policy Gradient Theorem \cite{sutton2000policy} available for RL with cumulative rewards, we derive a new Variational Policy Gradient Theorem for RL with general utilities, which establishes that the gradient may be obtained as the solution of a stochastic saddle point problem involving the Fenchel dual of the utility function. We develop a variational Monte Carlo gradient estimation algorithm to compute the policy gradient based on sample paths. Further, we prove that the variational policy gradient scheme converges globally to the optimal policy for the general objective, and we also establish its rate of convergence that matches or improves the convergence rate available in the case of RL with cumulative rewards.

[Reverse-engineering recurrent neural network solutions to a hierarchical inference task for mice](#)

- Rylan Schaeffer, Mikail Khona, Leenoy Meshulam, Brain Laboratory International, Ila Fiete
- abstract@[open-review](#): We study how recurrent neural networks (RNNs) solve a hierarchical inference task involving two latent variables and disparate timescales separated by 1-2 orders of magnitude. The task is of interest to the International Brain Laboratory, a global collaboration of experimental and theoretical neuroscientists studying how the mammalian brain generates behavior. We make four discoveries. First, RNNs learn behavior that is quantitatively similar to ideal Bayesian baselines. Second, RNNs perform inference by learning a two-dimensional subspace defining beliefs about the latent variables. Third, the geometry of RNN dynamics reflects an induced coupling between the two separate inference processes necessary to solve the task. Fourth, we perform model compression through a novel form of knowledge distillation on hidden representations -- Representations and Dynamics Distillation (RADD)-- to reduce the RNN dynamics to a low-dimensional, highly interpretable model. This technique promises a useful tool for interpretability of high dimensional nonlinear dynamical systems. Altogether, this work yields predictions to guide exploration and analysis of mouse neural data and circuitry.

[Temporal Positive-unlabeled Learning for Biomedical Hypothesis Generation via Risk Estimation](#)

- Uchenna Akujuobi, Jun Chen, Mohamed Elhoseiny, Michael Spranger, Xiangliang Zhang
- abstract@[open-review](#): Understanding the relationships between biomedical terms like viruses, drugs, and symptoms is essential in the fight against diseases. Many attempts have been made to introduce the use of machine learning to the scientific process of hypothesis generation (HG), which refers to the discovery of meaningful implicit connections between biomedical terms. However, most existing methods fail to truly capture the temporal dynamics of scientific term relations and also assume unobserved connections to be irrelevant (i.e., in a positive-negative (PN) learning setting). To break these limits, we formulate this HG problem as future connectivity prediction task on a dynamic attributed graph via positive-unlabeled (PU) learning. Then, the key is to capture the temporal evolution of node pair (term pair) relations from just the positive and unlabeled data. We propose a variational inference model to estimate the positive prior, and incorporate it in the learning of node pair embeddings, which are then used for link prediction. Experiment results on real-world biomedical term relationship datasets and case study analyses on a COVID-19 dataset validate the effectiveness of the proposed model.

[Efficient Low Rank Gaussian Variational Inference for Neural Networks](#)

- Marcin Tomczak, Siddharth Swaroop, Richard Turner
- abstract@[open-review](#): Bayesian neural networks are enjoying a renaissance driven in part by recent advances in variational inference (VI). The most common form of VI employs a fully factorized or mean-field distribution, but this is known to suffer from several pathologies, especially as we expect posterior distributions with highly correlated parameters. Current algorithms that capture these correlations with a Gaussian approximating family are difficult to scale to large models due to computational costs and high variance of gradient updates. By using a new form of the reparametrization trick, we derive a computationally efficient algorithm for performing VI with a Gaussian family with a low-rank plus diagonal covariance structure. We scale to deep feed-forward and convolutional architectures. We find that adding low-rank terms to parametrized diagonal covariance does not improve predictive

performance except on small networks, but low-rank terms added to a constant diagonal covariance improves performance on small and large-scale network architectures.

[Privacy Amplification via Random Check-Ins](#)

- Borja Balle, Peter Kairouz, Brendan McMahan, Om Thakkar, Abhradeep Guha Thakurta
- abstract@[open-review](#): Differentially Private Stochastic Gradient Descent (DP-SGD) forms a fundamental building block in many applications for learning over sensitive data. Two standard approaches, privacy amplification by subsampling, and privacy amplification by shuffling, permit adding lower noise in DP-SGD than via naive schemes. A key assumption in both these approaches is that the elements in the data set can be uniformly sampled, or be uniformly permuted --- constraints that may become prohibitive when the data is processed in a decentralized or distributed fashion. In this paper, we focus on conducting iterative methods like DP-SGD in the setting of federated learning (FL) wherein the data is distributed among many devices (clients). Our main contribution is the random check-in distributed protocol, which crucially relies only on randomized participation decisions made locally and independently by each client. It has privacy/accuracy trade-offs similar to privacy amplification by subsampling/shuffling. However, our method does not require server-initiated communication, or even knowledge of the population size. To our knowledge, this is the first privacy amplification tailored for a distributed learning framework, and it may have broader applicability beyond FL. Along the way, we improve the privacy guarantees of amplification by shuffling and show that, in practical regimes, this improvement allows for similar privacy and utility using data from an order of magnitude fewer users.

[Probabilistic Circuits for Variational Inference in Discrete Graphical Models](#)

- Andy Shih, Stefano Ermon
- abstract@[open-review](#): Inference in discrete graphical models with variational methods is difficult because of the inability to re-parameterize gradients of the Evidence Lower Bound (ELBO). Many sampling-based methods have been proposed for estimating these gradients, but they suffer from high bias or variance. In this paper, we propose a new approach that leverages the tractability of probabilistic circuit models, such as Sum Product Networks (SPN), to compute ELBO gradients exactly (without sampling) for a certain class of densities. In particular, we show that selective-SPNs are suitable as an expressive variational distribution, and prove that when the log-density of the target model is a polynomial the corresponding ELBO can be computed analytically. To scale to graphical models with thousands of variables, we develop an efficient and effective construction of selective-SPNs with size $(O(kn))$, where (n) is the number of variables and (k) is an adjustable hyperparameter. We demonstrate our approach on three types of graphical models -- Ising models, Latent Dirichlet Allocation, and factor graphs from the UAI Inference Competition. Selective-SPNs give a better lower bound than mean-field and structured mean-field, and is competitive with approximations that do not provide a lower bound, such as Loopy Belief Propagation and Tree-Reweighted Belief Propagation. Our results show that probabilistic circuits are promising tools for variational inference in discrete graphical models as they combine tractability and expressivity.

[Your Classifier can Secretly Suffice Multi-Source Domain Adaptation](#)

- Naveen Venkat, Jogendra Nath Kundu, Durgesh Singh, Ambareesh Revanur, Venkatesh Babu R
- abstract@[open-review](#): Multi-Source Domain Adaptation (MSDA) deals with the transfer of task knowledge from multiple labeled source domains to an unlabeled target domain, under a domain-shift. Existing methods aim to minimize this domain-shift using auxiliary distribution alignment objectives. In this work, we present a different perspective to MSDA wherein deep models are observed to implicitly align the domains under label supervision. Thus, we aim to utilize implicit alignment without additional training objectives to perform adaptation. To this end, we use pseudo-labeled target samples and enforce a classifier agreement on the pseudo-labels, a process called Self-supervised Implicit Alignment (SImpAl). We find that SImpAl readily works even under category-shift among the source domains. Further, we propose classifier agreement as a cue to determine the training convergence, resulting in a simple training algorithm. We provide a thorough evaluation of our approach on five benchmarks, along with detailed insights into each component of our approach.

[Labelling unlabelled videos from scratch with multi-modal self-supervision](#)

- Yuki Asano, Mandela Patrick, Christian Rupprecht, Andrea Vedaldi
- abstract@[open-review](#): A large part of the current success of deep learning lies in the effectiveness of data -- more precisely: of labeled data. Yet, labelling a dataset with human annotation continues to carry high costs, especially for videos. While in the image domain, recent methods have allowed to generate meaningful (pseudo-) labels for unlabelled datasets without supervision, this development is missing for the video domain where learning feature representations is the current focus. In this work, we a) show that unsupervised labelling of a video dataset does not come for free from strong feature encoders and b) propose a novel clustering method that allows pseudo-labelling of a video dataset without any human annotations, by leveraging the natural correspondence between audio and visual modalities. An extensive analysis shows that the resulting clusters have high semantic overlap to ground truth human labels. We further introduce the first benchmarking results on unsupervised labelling of common video datasets.

[A Non-Asymptotic Analysis for Stein Variational Gradient Descent](#)

- Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, Arthur Gretton
- abstract@[open-review](#): We study the Stein Variational Gradient Descent (SVGD) algorithm, which optimises a set of particles to approximate a target probability distribution $\pi^{\text{proto}} e^{-V}$ on R^d . In the population limit, SVGD performs gradient descent in the space of probability distributions on the KL divergence with respect to π , where the gradient is smoothed through a kernel integral operator. In this paper, we provide a novel finite time analysis for the SVGD algorithm. We provide a descent lemma establishing that the algorithm decreases the objective at each iteration, and rates of convergence. We also provide a convergence result of the finite particle system corresponding to the practical implementation of SVGD to its population version.

[Robust Meta-learning for Mixed Linear Regression with Small Batches](#)

- Weihao Kong, Raghav Somani, Sham Kakade, Sewoong Oh
- abstract@[open-review](#): A common challenge faced in practical supervised learning, such as medical image processing and robotic interactions, is that there are plenty of tasks but each task cannot afford to collect enough labeled examples to be learned in isolation. However, by exploiting the similarities across those tasks, one can hope to overcome such data scarcity. Under a canonical scenario where each task is drawn from a mixture of k linear regressions, we study a fundamental question: can abundant small-data tasks compensate for the lack of big-data tasks? Existing second moment based approaches of \cite{2020arXiv200208936K} show that such a trade-off is efficiently achievable, with the help of medium-sized tasks with $\Omega(k^{1/2})$ examples each. However, this algorithm is brittle in two important scenarios. The predictions can be arbitrarily bad (i) even with only a few outliers in the dataset; or (ii) even if the medium-sized tasks are slightly smaller with $\Theta(k^{1/2})$ examples each. We introduce a spectral approach that is simultaneously robust under both scenarios. To this end, we first design a novel outlier-robust principal component analysis algorithm that achieves an optimal accuracy. This is followed by a sum-of-squares algorithm to exploit the information from higher order moments. Together, this approach is robust against outliers and achieves a graceful statistical trade-off; the lack of $\Omega(k^{1/2})$ -size tasks can be compensated for with smaller tasks, which can now be as small as $\mathcal{O}(\log k)$.

[Bayesian Deep Learning and a Probabilistic Perspective of Generalization](#)

- Andrew G. Wilson, Pavel Izmailov
- abstract@[open-review](#): The key distinguishing property of a Bayesian approach is marginalization, rather than using a single setting of weights. Bayesian marginalization can particularly improve the accuracy and calibration of modern deep neural networks, which are typically underspecified by the data, and can represent many compelling but different solutions. We show that deep ensembles provide an effective mechanism for approximate Bayesian marginalization, and propose a related approach that further improves the predictive distribution by marginalizing within basins of attraction, without significant overhead. We also investigate the prior over functions implied by a vague distribution over neural network weights, explaining the generalization properties of such models from a probabilistic perspective. From this perspective, we explain results that have been presented as mysterious and distinct to neural network generalization, such as the ability to fit images with random labels, and show that these results can be reproduced with Gaussian processes. We also show that Bayesian model averaging alleviates double descent, resulting in monotonic performance improvements with increased flexibility.

[Unsupervised Learning of Object Landmarks via Self-Training Correspondence](#)

- Dimitrios Mallis, Enrique Sanchez, Matthew Bell, Georgios Tzimiropoulos
- abstract@[open-review](#): This paper addresses the problem of unsupervised discovery of object landmarks. We take a different path compared to that of existing works, based on 2 novel perspectives: (1) Self-training: starting from generic keypoints, we propose a self-training approach where the goal is to learn a detector that improves itself becoming more and more tuned to object landmarks. (2) Correspondence: we identify correspondence as a key objective for unsupervised landmark discovery and propose an optimization scheme which alternates between recovering object landmark correspondence across different images via clustering and learning an object landmark descriptor without labels. Compared to previous works, our approach can learn landmarks that are more flexible in terms of capturing large changes in viewpoint. We show the favourable properties of our method on a variety of difficult datasets including LS3D, BBCPose and Human3.6M. Code is available at <https://github.com/malldimi1/UnsupervisedLandmarks>

[Randomized tests for high-dimensional regression: A more efficient and powerful solution](#)

- Yue Li, Ilmun Kim, Yuting Wei
- abstract@[open-review](#): We investigate the problem of testing the global null in the high-dimensional regression models when the feature dimension p grows proportionally to the number of observations n . Despite a number of prior work studying this problem, whether there exists a test that is model-agnostic, efficient to compute and enjoys a high power, still remains unsettled. In this paper, we answer this question in the affirmative by leveraging the random projection techniques, and propose a testing procedure that blends the classical F -test with a random projection step. When combined with a systematic choice of the projection dimension, the proposed procedure is proved to be minimax optimal and, meanwhile, reduces the computation and data storage requirements. We illustrate our results in various scenarios when the underlying feature matrix exhibits an intrinsic lower dimensional structure (such as approximate low-rank or has exponential/polynomial eigen-decay), and it turns out that the proposed test achieves sharp adaptive rates. Our theoretical findings are further validated by comparisons to other state-of-the-art tests on synthetic data.

[Learning Representations from Audio-Visual Spatial Alignment](#)

- Pedro Morgado, Yi Li, Nuno Vasconcelos
- abstract@[open-review](#): We introduce a novel self-supervised pretext task for learning representations from audio-visual content. Prior work on audio-visual representation learning leverages correspondences at the video level. Approaches based on audio-visual correspondence (AVC) predict whether audio and video clips originate from the same or different video instances. Audio-visual temporal synchronization (AVTS) further discriminates negative pairs originated from the same video instance but at different moments in time. While these approaches learn high-quality representations for downstream tasks such as action recognition, they completely disregard the spatial cues of audio and visual signals naturally occurring in the real world. To learn from these spatial cues, we tasked a network to perform contrastive audio-visual spatial alignment of 360-degree video and spatial audio. The ability to perform spatial alignment is enhanced by reasoning over the full spatial content of the 360-degree video using a transformer architecture to combine representations from multiple viewpoints. The advantages of the proposed pretext task are demonstrated on a variety of audio and visual downstream tasks, including audio-visual correspondence, spatial alignment, action recognition and video semantic segmentation. Dataset and code are available at <https://github.com/pedro-morgado/AVSpatialAlignment>.

[Generative View Synthesis: From Single-view Semantics to Novel-view Images](#)

- Tewodros Amberbir Habtegebrial, Varun Jampani, Orazio Gallo, Didier Stricker
- abstract@[open-review](#): Content creation, central to applications such as virtual reality, can be tedious and time-consuming. Recent image synthesis methods simplify this task by offering tools to generate new views from as little as a single input image, or by converting a semantic map into a photorealistic image. We propose to push the envelope further, and introduce Generative View Synthesis (GVS) that can synthesize multiple photorealistic views of a scene given a single semantic map. We show that the sequential application of existing techniques, e.g., semantics-to-image translation followed by monocular view synthesis, fail at capturing the scene's structure. In contrast, we solve the semantics-to-image translation in concert with the estimation of the 3D layout of the scene, thus producing geometrically consistent novel views that preserve semantic structures. We first lift the input 2D semantic map onto a 3D layered representation of the scene in feature space, thereby preserving the semantic labels of 3D geometric structures. We then project the layered features onto the target views to generate the final novel-view images. We verify the strengths of our method and compare it with several advanced baselines on three different datasets. Our approach also allows for style manipulation and image editing operations, such as the addition or removal of objects, with simple manipulations of the input style images and semantic maps respectively. For code and additional results, visit the project page at <https://gvsnet.github.io>

[Towards More Practical Adversarial Attacks on Graph Neural Networks](#)

- Jiaqi Ma, Shuangrui Ding, Qiaozhu Mei
- abstract@[open-review](#): We study the black-box attacks on graph neural networks (GNNs) under a novel and realistic constraint: attackers have access to only a subset of nodes in the network, and they can only attack a small number of them. A node selection step is essential under this setup. We demonstrate that the structural inductive biases of GNN models can be an effective source for this type of attacks. Specifically, by exploiting the connection between the backward propagation of GNNs and random walks, we show that the common gradient-based white-box attacks can be generalized to the black-box setting via the connection between the gradient and an importance score similar to PageRank. In practice, we find attacks based on this importance score indeed increase the classification loss by a large margin, but they fail to significantly increase the mis-classification rate. Our theoretical and empirical analyses suggest that there is a discrepancy between the loss and mis-classification rate, as the latter presents a diminishing-return pattern when the number of attacked nodes increases. Therefore, we propose a greedy procedure to correct the importance score that takes into account of the diminishing-return pattern. Experimental results show that the proposed procedure can significantly increase the mis-classification rate of common GNNs on real-world data without access to model parameters nor predictions.

[Multi-Task Reinforcement Learning with Soft Modularization](#)

- Ruihan Yang, Huazhe Xu, YI WU, Xiaolong Wang
- abstract@[open-review](#): Multi-task learning is a very challenging problem in reinforcement learning. While training multiple tasks jointly allow the policies to share parameters across different tasks, the optimization problem becomes non-trivial: It remains unclear what parameters in the network

should be reused across tasks, and how the gradients from different tasks may interfere with each other. Thus, instead of naively sharing parameters across tasks, we introduce an explicit modularization technique on policy representation to alleviate this optimization issue. Given a base policy network, we design a routing network which estimates different routing strategies to reconfigure the base network for each task. Instead of directly selecting routes for each task, our task-specific policy uses a method called soft modularization to softly combine all the possible routes, which makes it suitable for sequential tasks. We experiment with various robotics manipulation tasks in simulation and show our method improves both sample efficiency and performance over strong baselines by a large margin.

[Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models](#)

- Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, Tom Claassen
- abstract@[open-review](#): Shapley values underlie one of the most popular model-agnostic methods within explainable artificial intelligence. These values are designed to attribute the difference between a model's prediction and an average baseline to the different features used as input to the model. Being based on solid game-theoretic principles, Shapley values uniquely satisfy several desirable properties, which is why they are increasingly used to explain the predictions of possibly complex and highly non-linear machine learning models. Shapley values are well calibrated to a user's intuition when features are independent, but may lead to undesirable, counterintuitive explanations when the independence assumption is violated. In this paper, we propose a novel framework for computing Shapley values that generalizes recent work that aims to circumvent the independence assumption. By employing Pearl's do-calculus, we show how these 'causal' Shapley values can be derived for general causal graphs without sacrificing any of their desirable properties. Moreover, causal Shapley values enable us to separate the contribution of direct and indirect effects. We provide a practical implementation for computing causal Shapley values based on causal chain graphs when only partial information is available and illustrate their utility on a real-world example.

[On the training dynamics of deep networks with \$L_2\$ regularization](#)

- Aitor Lewkowycz, Guy Gur-Ari
- abstract@[open-review](#): We study the role of L_2 regularization in deep learning, and uncover simple relations between the performance of the model, the L_2 coefficient, the learning rate, and the number of training steps. These empirical relations hold when the network is overparameterized. They can be used to predict the optimal regularization parameter of a given model. In addition, based on these observations we propose a dynamical schedule for the regularization parameter that improves performance and speeds up training. We test these proposals in modern image classification settings. Finally, we show that these empirical relations can be understood theoretically in the context of infinitely wide networks. We derive the gradient flow dynamics of such networks, and compare the role of L_2 regularization in this context with that of linear models.

[Improved Algorithms for Convex-Concave Minimax Optimization](#)

- Yuanhao Wang, Jian Li
- abstract@[open-review](#): This paper studies minimax optimization problems $\min_x \max_y f(x, y)$, where $f(x, y)$ is m_x -strongly convex with respect to x , m_y -strongly concave with respect to y and $(L_x, L_{\{x\}y}, L_y)$ -smooth. Zhang et al. [cite{zhang2019lower}](#) provided the following lower bound of the gradient complexity for any first-order method: $\Omega(\sqrt{\frac{L_x}{m_x}} + \frac{L_{\{x\}y}}{m_x})$. This paper proposes a new algorithm and proved a gradient complexity bound of $\tilde{O}(\sqrt{\frac{L_x}{m_x}} + \frac{L_{\{x\}y}}{m_x} + \frac{L_{\{x\}y}}{m_y}) \ln(1/\epsilon) \Big/ \sqrt{\frac{L_x}{m_x} + \frac{L_{\{x\}y}}{m_y}}$, where $L = \max\{L_x, L_{\{x\}y}, L_y\}$. This improves over the best known upper bound $\tilde{O}(\sqrt{\frac{L^2}{m_x m_y}} \ln^3(1/\epsilon))$ by Lin et al. [cite{lin2020near}](#). Our bound achieves linear convergence rate and tighter dependency on condition numbers, especially when $L_{\{x\}y} \ll L$ (i.e., the weak interaction regime). Via simple reduction, our new bound also implies improved bounds for strongly convex-concave problems and convex-concave problems. When f is quadratic, we can further improve the bound to $O(\sqrt{\frac{L^2}{m_x m_y}} + o(1) \ln(1/\epsilon))$, which matches the lower bound up to a sub-polynomial factor.

[Deep Variational Instance Segmentation](#)

- Jialin Yuan, Chao Chen, Fuxin Li
- abstract@[open-review](#): Instance segmentation, which seeks to obtain both class and instance labels for each pixel in the input image, is a challenging task in computer vision. State-of-the-art algorithms often employ a search-based strategy, which first divides the output image with a regular grid and generate proposals at each grid cell, then the proposals are classified and boundaries refined. In this paper, we propose a novel algorithm that directly utilizes a fully convolutional network (FCN) to predict instance labels. Specifically, we propose a variational relaxation of instance segmentation as minimizing an optimization functional for a piecewise-constant segmentation problem, which can be used to train an FCN end-to-end. It extends the classical Mumford-Shah variational segmentation algorithm to be able to handle the permutation-invariant ground truth in instance segmentation. Experiments on PASCAL VOC 2012 and the MSCOCO 2017 dataset show that the proposed approach efficiently tackles the instance segmentation task.

[Learning Implicit Functions for Topology-Varying Dense 3D Shape Correspondence](#)

- Feng Liu, Xiaoming Liu
- abstract@[open-review](#): The goal of this paper is to learn dense 3D shape correspondence for topology-varying objects in an unsupervised manner. Conventional implicit functions estimate the occupancy of a 3D point given a shape latent code. Instead, our novel implicit function produces a part embedding vector for each 3D point, which is assumed to be similar to its densely corresponded point in another 3D shape of the same object category. Furthermore, we implement dense correspondence through an inverse function mapping from the part embedding to a corresponded 3D point. Both functions are jointly learned with several effective loss functions to realize our assumption, together with the encoder generating the shape latent code. During inference, if a user selects an arbitrary point on the source shape, our algorithm can automatically generate a confidence score indicating whether there is a correspondence on the target shape, as well as the corresponding semantic point if there is. Such a mechanism inherently benefits man-made objects with different part constitutions. The effectiveness of our approach is demonstrated through unsupervised 3D semantic correspondence and shape segmentation.

[Deep Multimodal Fusion by Channel Exchanging](#)

- Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, Junzhou Huang
- abstract@[open-review](#): Deep multimodal fusion by using multiple sources of data for classification or regression has exhibited a clear advantage over the unimodal counterpart on various applications. Yet, current methods including aggregation-based and alignment-based fusion are still inadequate in balancing the trade-off between inter-modal fusion and intra-modal processing, incurring a bottleneck of performance improvement. To this end, this paper proposes Channel-Exchanging-Network (CEN), a parameter-free multimodal fusion framework that dynamically exchanges channels between sub-networks of different modalities. Specifically, the channel exchanging process is self-guided by individual channel importance that is measured by the magnitude of Batch-Normalization (BN) scaling factor during training. The validity of such exchanging process is also guaranteed by sharing convolutional filters yet keeping separate BN layers across modalities, which, as an add-on benefit, allows our multimodal architecture to be almost as compact as a unimodal network. Extensive experiments on semantic segmentation via RGB-D data and image translation through multi-domain input

verify the effectiveness of our CEN compared to current state-of-the-art methods. Detailed ablation studies have also been carried out, which provably affirm the advantage of each component we propose. Our code is available at <https://github.com/yikaiw/CEN>.

[Hierarchically Organized Latent Modules for Exploratory Search in Morphogenetic Systems](#)

- Mayalen Etcheverry, Clément Moulin-Frier, Pierre-Yves Oudeyer
- abstract@[open-review](#): Self-organization of complex morphological patterns from local interactions is a fascinating phenomenon in many natural and artificial systems. In the artificial world, typical examples of such morphogenetic systems are cellular automata. Yet, their mechanisms are often very hard to grasp and so far scientific discoveries of novel patterns have primarily been relying on manual tuning and ad hoc exploratory search. The problem of automated diversity-driven discovery in these systems was recently introduced [26, 62], highlighting that two key ingredients are autonomous exploration and unsupervised representation learning to describe the degrees of variations in the patterns. In this paper, we motivate the need for what we call Meta-diversity search, arguing that there is not a unique ground truth interesting diversity as it strongly depends on the final observer and its motives. Using a continuous game-of-life system for experiments, we provide empirical evidences that relying on monolithic architectures for the behavioral embedding design tends to bias the final discoveries (both for hand-defined and unsupervisedly-learned features) which are unlikely to be aligned with the interest of a final end-user. To address these issues, we introduce a novel dynamic and modular architecture that enables unsupervised learning of a hierarchy of diverse representations. Combined with intrinsically motivated goal exploration algorithms, we show that this system forms a discovery assistant that can efficiently adapt its diversity search towards preferences of a user using only a very small amount of user feedback.

[AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity](#)

- Silviu-Marian Udrescu, Andrew Tan, Jiahai Feng, Orisvaldo Neto, Tailin Wu, Max Tegmark
- abstract@[open-review](#): We present an improved method for symbolic regression that seeks to fit data to formulas that are Pareto-optimal, in the sense of having the best accuracy for a given complexity. It improves on the previous state-of-the-art by typically being orders of magnitude more robust toward noise and bad data, and also by discovering many formulas that stumped previous methods. We develop a method for discovering generalized symmetries (arbitrary modularity in the computational graph of a formula) from gradient properties of a neural network fit. We use normalizing flows to generalize our symbolic regression method to probability distributions from which we only have samples, and employ statistical hypothesis testing to accelerate robust brute-force search.

[Delay and Cooperation in Nonstochastic Linear Bandits](#)

- Shinji Ito, Daisuke Hatano, Hanna Sumita, Kei Takemura, Takuro Fukunaga, Naonori Kakimura, Ken-Ichi Kawarabayashi
- abstract@[open-review](#): This paper offers a nearly optimal algorithm for online linear optimization with delayed bandit feedback. Online linear optimization with bandit feedback, or nonstochastic linear bandits, provides a generic framework for sequential decision-making problems with limited information. This framework, however, assumes that feedback can be observed just after choosing the action, and, hence, does not apply directly to many practical applications, in which the feedback can often only be obtained after a while. To cope with such situations, we consider problem settings in which the feedback can be observed $\$d\$$ rounds after the choice of an action, and propose an algorithm for which the expected regret is $\$tilde{O}(\sqrt{m(m+d)T})$, ignoring logarithmic factors in m and T , where m and T denote the dimensionality of the action set and the number of rounds, respectively. This algorithm achieves nearly optimal performance, as we are able to show that arbitrary algorithms suffer the regret of $\$Omega(\sqrt{m(m+d)T})$ in the worst case. To develop the algorithm, we introduce a technique we refer to as distribution truncation, which plays an essential role in bounding the regret. We also apply our approach to cooperative bandits, as studied by Cesa-Bianchi et al. [17] and Bar-On and Mansour [12], and extend their results to the linear bandits setting.

[Probabilistic Orientation Estimation with Matrix Fisher Distributions](#)

- David Mohlin, Josephine Sullivan, Gárald Bianchi
- abstract@[open-review](#): This paper focuses on estimating probability distributions over the set of 3D rotations ($SO(3)$) using deep neural networks. Learning to regress models to the set of rotations is inherently difficult due to differences in topology between R^N and $SO(3)$. We overcome this issue by using a neural network to output the parameters for a matrix Fisher distribution since these parameters are homeomorphic to R^9 . By using a negative log likelihood loss for this distribution we get a loss which is convex with respect to the network outputs. By optimizing this loss we improve state-of-the-art on several challenging applicable datasets, namely Pascal3D+, ModelNet10-SO(3). Our code is available at https://github.com/Davmo049/Publicproborientationestimationwithmatrix_fisherdistribution

[Minimax Dynamics of Optimally Balanced Spiking Networks of Excitatory and Inhibitory Neurons](#)

- Qianyi Li, Cengiz Pehlevan
- abstract@[open-review](#): Excitation-inhibition balance is ubiquitously observed in the cortex. Recent studies suggest an intriguing link between balance on fast timescales, tight balance, and efficient information coding with spikes. We further this connection by taking a principled approach to optimal balanced networks of excitatory (E) and inhibitory(I) neurons. By deriving E-I spiking neural networks from greedy spike-based optimizations of constrained minimax objectives, we show that tight balance arises from correcting for deviations from the minimax optimum. We predict specific neuron firing rates in the networks by solving the minimax problems, going beyond statistical theories of balanced networks. We design minimax objectives for reconstruction of an input signal, associative memory, and storage of manifold attractors, and derive from them E-I networks that perform the computation. Overall, we present a novel normative modeling approach for spiking E-I networks, going beyond the widely-used energy-minimizing networks that violate Dale's law. Our networks can be used to model cortical circuits and computations.

[Telescoping Density-Ratio Estimation](#)

- Benjamin Rhodes, Kai Xu, Michael U. Gutmann
- abstract@[open-review](#): Density-ratio estimation via classification is a cornerstone of unsupervised learning. It has provided the foundation for state-of-the-art methods in representation learning and generative modelling, with the number of use-cases continuing to proliferate. However, it suffers from a critical limitation: it fails to accurately estimate ratios p/q for which the two densities differ significantly. Empirically, we find this occurs whenever the KL divergence between p and q exceeds tens of nats. To resolve this limitation, we introduce a new framework, telescoping density-ratio estimation (TRE), that enables the estimation of ratios between highly dissimilar densities in high-dimensional spaces. Our experiments demonstrate that TRE can yield substantial improvements over existing single-ratio methods for mutual information estimation, representation learning and energy-based modelling.

[Towards Deeper Graph Neural Networks with Differentiable Group Normalization](#)

- Kaixiong Zhou, Xiao Huang, Yuening Li, Daochen Zha, Rui Chen, Xia Hu
- abstract@[open-review](#): Graph neural networks (GNNs), which learn the representation of a node by aggregating its neighbors, have become an effective computational tool in downstream applications. Over-smoothing is one of the key issues which limit the performance of GNNs as the number of layers increases. It is because the stacked aggregators would make node representations converge to indistinguishable vectors. Several attempts have been made to tackle the issue by bringing linked node pairs close and unlinked pairs distinct. However, they often ignore the intrinsic community structures and

would result in sub-optimal performance. The representations of nodes within the same community/class need be similar to facilitate the classification, while different classes are expected to be separated in embedding space. To bridge the gap, we introduce two over-smoothing metrics and a novel technique, i.e., differentiable group normalization (DGN). It normalizes nodes within the same group independently to increase their smoothness, and separates node distributions among different groups to significantly alleviate the over-smoothing issue. Experiments on real-world datasets demonstrate that DGN makes GNN models more robust to over-smoothing and achieves better performance with deeper GNNs.

[Stochastic Optimization for Performative Prediction](#)

- Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, Moritz Hardt
- abstract@[open-review](#): In performative prediction, the choice of a model influences the distribution of future data, typically through actions taken based on the model's predictions. We initiate the study of stochastic optimization for performative prediction. What sets this setting apart from traditional stochastic optimization is the difference between merely updating model parameters and deploying the new model. The latter triggers a shift in the distribution that affects future data, while the former keeps the distribution as is. Assuming smoothness and strong convexity, we prove rates of convergence for both greedily deploying models after each stochastic update (greedy deploy) as well as for taking several updates before redeploying (lazy deploy). In both cases, our bounds smoothly recover the optimal $\mathcal{O}(1/k)$ rate as the strength of performativity decreases. Furthermore, they illustrate how depending on the strength of performative effects, there exists a regime where either approach outperforms the other. We experimentally explore the trade-off on both synthetic data and a strategic classification simulator.

[Learning Differentiable Programs with Admissible Neural Heuristics](#)

- Ameesh Shah, Eric Zhan, Jennifer Sun, Abhinav Verma, Yisong Yue, Swarat Chaudhuri
- abstract@[open-review](#): We study the problem of learning differentiable functions expressed as programs in a domain-specific language. Such programmatic models can offer benefits such as composability and interpretability; however, learning them requires optimizing over a combinatorial space of program "architectures". We frame this optimization problem as a search in a weighted graph whose paths encode top-down derivations of program syntax. Our key innovation is to view various classes of neural networks as continuous relaxations over the space of programs, which can then be used to complete any partial program. All the parameters of this relaxed program can be trained end-to-end, and the resulting training loss is an approximately admissible heuristic that can guide the combinatorial search. We instantiate our approach on top of the A* and Iterative Deepening Depth-First Search algorithms and use these algorithms to learn programmatic classifiers in three sequence classification tasks. Our experiments show that the algorithms outperform state-of-the-art methods for program learning, and that they discover programmatic classifiers that yield natural interpretations and achieve competitive accuracy.

[Improved guarantees and a multiple-descent curve for Column Subset Selection and the Nystrom method](#)

- Michal Derezhinski, Rajiv Khanna, Michael W. Mahoney
- abstract@[open-review](#): The Column Subset Selection Problem (CSSP) and the Nystrom method are among the leading tools for constructing small low-rank approximations of large datasets in machine learning and scientific computing. A fundamental question in this area is: how well can a data subset of size k compete with the best rank k approximation? We develop techniques which exploit spectral properties of the data matrix to obtain improved approximation guarantees which go beyond the standard worst-case analysis. Our approach leads to significantly better bounds for datasets with known rates of singular value decay, e.g., polynomial or exponential decay. Our analysis also reveals an intriguing phenomenon: the approximation factor as a function of k may exhibit multiple peaks and valleys, which we call a multiple-descent curve. A lower bound we establish shows that this behavior is not an artifact of our analysis, but rather it is an inherent property of the CSSP and Nystrom tasks. Finally, using the example of a radial basis function (RBF) kernel, we show that both our improved bounds and the multiple-descent curve can be observed on real datasets simply by varying the RBF parameter.

[Domain Adaptation as a Problem of Inference on Graphical Models](#)

- Kun Zhang, Mingming Gong, Petar Stojanov, Biwei Huang, QINGSONG LIU, Clark Glymour
- abstract@[open-review](#): This paper is concerned with data-driven unsupervised domain adaptation, where it is unknown in advance how the joint distribution changes across domains, i.e., what factors or modules of the data distribution remain invariant or change across domains. To develop an automated way of domain adaptation with multiple source domains, we propose to use a graphical model as a compact way to encode the change property of the joint distribution, which can be learned from data, and then view domain adaptation as a problem of Bayesian inference on the graphical models. Such a graphical model distinguishes between constant and varied modules of the distribution and specifies the properties of the changes across domains, which serves as prior knowledge of the changing modules for the purpose of deriving the posterior of the target variable Y in the target domain. This provides an end-to-end framework of domain adaptation, in which additional knowledge about how the joint distribution changes, if available, can be directly incorporated to improve the graphical representation. We discuss how causality-based domain adaptation can be put under this umbrella. Experimental results on both synthetic and real data demonstrate the efficacy of the proposed framework for domain adaptation.

[Network size and size of the weights in memorization with two-layers neural networks](#)

- Sébastien Bubeck, Ronen Eldan, Yin Tat Lee, Dan Mikulincer
- abstract@[open-review](#): In 1988, Eric B. Baum showed that two-layers neural networks with threshold activation function can perfectly memorize the binary labels of n points in general position in R^d using only $\lceil n/d \rceil$ neurons. We observe that with ReLU networks, using four times as many neurons one can fit arbitrary real labels. Moreover, for approximate memorization up to error ϵ , the neural tangent kernel can also memorize with only $O(\left(\frac{n}{d}\right) \cdot \log(1/\epsilon))$ neurons (assuming that the data is well dispersed too). We show however that these constructions give rise to networks where the magnitude of the neurons' weights are far from optimal. In contrast we propose a new training procedure for ReLU networks, based on a complex recombination of the neurons, for which we show approximate memorization with both $O(\left(\frac{n}{d}\right) \cdot \frac{\log(1/\epsilon)}{\epsilon})$ neurons, as well as nearly-optimal size of the weights.

[Certifying Strategyproof Auction Networks](#)

- Michael Curry, Ping-yeh Chiang, Tom Goldstein, John Dickerson
- abstract@[open-review](#): Optimal auctions maximize a seller's expected revenue subject to individual rationality and strategyproofness for the buyers. Myerson's seminal work in 1981 settled the case of auctioning a single item; however, subsequent decades of work have yielded little progress moving beyond a single item, leaving the design of revenue-maximizing auctions as a central open problem in the field of mechanism design. A recent thread of work in ``differentiable economics'' has used tools from modern deep learning to instead learn good mechanisms. We focus on the RegretNet architecture, which can represent auctions with arbitrary numbers of items and participants; it is trained to be empirically strategyproof, but the property is never exactly verified leaving potential loopholes for market participants to exploit. We propose ways to explicitly verify strategyproofness under a particular valuation profile using techniques from the neural network verification literature. Doing so requires making several modifications to the RegretNet architecture in order to represent it exactly in an integer program. We train our network and produce certificates in several settings, including settings for which the optimal strategyproof mechanism is not known.

[Continual Learning of Control Primitives : Skill Discovery via Reset-Games](#)

- Kelvin Xu, Siddharth Verma, Chelsea Finn, Sergey Levine
- abstract@[open-review](#): Reinforcement learning has the potential to automate the acquisition of behavior in complex settings, but in order for it to be successfully deployed, a number of practical challenges must be addressed. First, in real world settings, when an agent attempts a task and fails, the environment must somehow "reset" so that the agent can attempt the task again. While easy in simulation, this could require considerable human effort in the real world, especially if the number of trials is very large. Second, real world learning is often limited by challenges in exploration, as complex, temporally extended behavior is often times difficult to acquire with random exploration. In this work, we show how a single method can allow an agent to acquire skills with minimal supervision while removing the need for resets. We do this by exploiting the insight that the need to reset" an agent to a broad set of initial states for a learning task provides a natural setting to learn a diverse set of reset-skills." We propose a general-sum game formulation that naturally balances the objective of resetting and learning skills, and demonstrate that this approach improves performance on reset-free tasks, and additionally show that the skills we obtain can be used to significantly accelerate downstream learning.

[HOI Analysis: Integrating and Decomposing Human-Object Interaction](#)

- Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, Cewu Lu
- abstract@[open-review](#): Human-Object Interaction (HOI) consists of human, object and implicit interaction/verb. Different from previous methods that directly map pixels to HOI semantics, we propose a novel perspective for HOI learning in an analytical manner. In analogy to Harmonic Analysis, whose goal is to study how to represent the signals with the superposition of basic waves, we propose the HOI Analysis. We argue that coherent HOI can be decomposed into isolated human and object. Meanwhile, isolated human and object can also be integrated into coherent HOI again. Moreover, transformations between human-object pairs with the same HOI can also be easier approached with integration and decomposition. As a result, the implicit verb will be represented in the transformation function space. In light of this, we propose an Integration-Decomposition Network (IDN) to implement the above transformations and achieve state-of-the-art performance on widely-used HOI detection benchmarks. Code is available at [https://github.com/DirtyHarryLYL/HAKE-Action-Torch/tree>IDN-\(Integrating-Decomposing-Network\)](https://github.com/DirtyHarryLYL/HAKE-Action-Torch/tree>IDN-(Integrating-Decomposing-Network)).

[Strongly local p-norm-cut algorithms for semi-supervised learning and local graph clustering](#)

- Meng Liu, David F. Gleich
- abstract@[open-review](#): Graph based semi-supervised learning is the problem of learning a labeling function for the graph nodes given a few example nodes, often called seeds, usually under the assumption that the graph's edges indicate similarity of labels. This is closely related to the local graph clustering or community detection problem of finding a cluster or community of nodes around a given seed. For this problem, we propose a novel generalization of random walk, diffusion, or smooth function methods in the literature to a convex p-norm cut function. The need for our p-norm methods is that, in our study of existing methods, we find those principled methods based on eigenvector, spectral, random walk, or linear system often have difficulty capturing the correct boundary of a target label or target cluster. In contrast, 1-norm or maxflow-mincut based methods capture the boundary, but cannot grow from small seed set; hybrid procedures that use both have many hard to set parameters. In this paper, we propose a generalization of the objective function behind these methods involving p-norms. To solve the p-norm cut problem we give a strongly local algorithm -- one whose runtime depends on the size of the output rather than the size of the graph. Our method can be thought as a nonlinear generalization of the Anderson-Chung-Lang push procedure to approximate a personalized PageRank vector efficiently. Our procedure is general and can solve other types of nonlinear objective functions, such as p-norm variants of Huber losses. We provide a theoretical analysis of finding planted target clusters with our method and show that the p-norm cut functions improve on the standard Cheeger inequalities for random walk and spectral methods. Finally, we demonstrate the speed and accuracy of our new method in synthetic and real world datasets.

[Deep Direct Likelihood Knockoffs](#)

- Mukund Sudarshan, Wesley Tansey, Rajesh Ranganath
- abstract@[open-review](#): Predictive modeling often uses black box machine learning methods, such as deep neural networks, to achieve state-of-the-art performance. In scientific domains, the scientist often wishes to discover which features are actually important for making the predictions. These discoveries may lead to costly follow-up experiments and as such it is important that the error rate on discoveries is not too high. Model-X knockoffs enable important features to be discovered with control of the false discovery rate (FDR). However, knockoffs require rich generative models capable of accurately modeling the knockoff features while ensuring they obey the so-called "swap" property. We develop Deep Direct Likelihood Knockoffs (DDLK), which directly minimizes the KL divergence implied by the knockoff swap property. DDLK consists of two stages: it first maximizes the explicit likelihood of the features, then minimizes the KL divergence between the joint distribution of features and knockoffs and any swap between them. To ensure that the generated knockoffs are valid under any possible swap, DDLK uses the Gumbel-Softmax trick to optimize the knockoff generator under the worst-case swap. We find DDLK has higher power than baselines while controlling the false discovery rate on a variety of synthetic and real benchmarks including a task involving the largest COVID-19 health record dataset in the United States.

[Meta-Neighborhoods](#)

- Siyuan Shan, Yang Li, Junier B. Oliva
- abstract@[open-review](#): Making an adaptive prediction based on input is an important ability for general artificial intelligence. In this work, we step forward in this direction and propose a semi-parametric method, Meta-Neighborhoods, where predictions are made adaptively to the neighborhood of the input. We show that Meta-Neighborhoods is a generalization of k-nearest-neighbors. Due to the simpler manifold structure around a local neighborhood, Meta-Neighborhoods represent the predictive distribution $p(y | x)$ more accurately. To reduce memory and computation overheads, we propose induced neighborhoods that summarize the training data into a much smaller dictionary. A meta-learning based training mechanism is then exploited to jointly learn the induced neighborhoods and the model. Extensive studies demonstrate the superiority of our method.

[Neural Dynamic Policies for End-to-End Sensorimotor Learning](#)

- Shikhar Bahl, Mustafa Mukadam, Abhinav Gupta, Deepak Pathak
- abstract@[open-review](#): The current dominant paradigm in sensorimotor control, whether imitation or reinforcement learning, is to train policies directly in raw action spaces such as torque, joint angle, or end-effector position. This forces the agent to make decision at each point in training, and hence, limits the scalability to continuous, high-dimensional, and long-horizon tasks. In contrast, research in classical robotics has, for a long time, exploited dynamical systems as a policy representation to learn robot behaviors via demonstrations. These techniques, however, lack the flexibility and generalizability provided by deep learning or deep reinforcement learning and have remained under-explored in such settings. In this work, we begin to close this gap and embed dynamics structure into deep neural network-based policies by reparameterizing action spaces with differential equations. We propose Neural Dynamic Policies (NDPs) that make predictions in trajectory distribution space as opposed to prior policy learning methods where action represents the raw control space. The embedded structure allows us to perform end-to-end policy learning under both reinforcement and imitation learning setups. We show that NDPs achieve better or comparable performance to state-of-the-art approaches on many robotic control tasks using both reward-based training and demonstrations. Project video and code are available at: <https://shikharbahl.github.io/neural-dynamic-policies/>.

[A new inference approach for training shallow and deep generalized linear models of noisy interacting neurons](#)

- Gabriel Mahuas, Giulio Isacchini, Olivier Marre, Ulisse Ferrari, Thierry Mora

- abstract@[open-review](#): Generalized linear models are one of the most efficient paradigms for predicting the correlated stochastic activity of neuronal networks in response to external stimuli, with applications in many brain areas. However, when dealing with complex stimuli, the inferred coupling parameters often do not generalise across different stimulus statistics, leading to degraded performance and blowup instabilities. Here, we develop a two-step inference strategy that allows us to train robust generalised linear models of interacting neurons, by explicitly separating the effects of correlations in the stimulus from network interactions in each training step. Applying this approach to the responses of retinal ganglion cells to complex visual stimuli, we show that, compared to classical methods, the models trained in this way exhibit improved performance, are more stable, yield robust interaction networks, and generalise well across complex visual statistics. The method can be extended to deep convolutional neural networks, leading to models with high predictive accuracy for both the neuron firing rates and their correlations.

[Decision-Making with Auto-Encoding Variational Bayes](#)

- Romain Lopez, Pierre Boyeau, Nir Yosef, Michael Jordan, Jeffrey Regier
- abstract@[open-review](#): To make decisions based on a model fit with auto-encoding variational Bayes (AEVB), practitioners often let the variational distribution serve as a surrogate for the posterior distribution. This approach yields biased estimates of the expected risk, and therefore leads to poor decisions for two reasons. First, the model fit with AEVB may not equal the underlying data distribution. Second, the variational distribution may not equal the posterior distribution under the fitted model. We explore how fitting the variational distribution based on several objective functions other than the ELBO, while continuing to fit the generative model based on the ELBO, affects the quality of downstream decisions. For the probabilistic principal component analysis model, we investigate how importance sampling error, as well as the bias of the model parameter estimates, varies across several approximate posteriors when used as proposal distributions. Our theoretical results suggest that a posterior approximation distinct from the variational distribution should be used for making decisions. Motivated by these theoretical results, we propose learning several approximate proposals for the best model and combining them using multiple importance sampling for decision-making. In addition to toy examples, we present a full-fledged case study of single-cell RNA sequencing. In this challenging instance of multiple hypothesis testing, our proposed approach surpasses the current state of the art.

[Attribution Preservation in Network Compression for Reliable Network Interpretation](#)

- Geondo Park, June Yong Yang, Sung Ju Hwang, Eunho Yang
- abstract@[open-review](#): Neural networks embedded in safety-sensitive applications such as self-driving cars and wearable health monitors rely on two important techniques: input attribution for hindsight analysis and network compression to reduce its size for edge-computing. In this paper, we show that these seemingly unrelated techniques conflict with each other as network compression deforms the produced attributions, which could lead to dire consequences for mission-critical applications. This phenomenon arises due to the fact that conventional network compression methods only preserve the predictions of the network while ignoring the quality of the attributions. To combat the attribution inconsistency problem, we present a framework that can preserve the attributions while compressing a network. By employing the Weighted Collapsed Attribution Matching regularizer, we match the attribution maps of the network being compressed to its pre-compression former self. We demonstrate the effectiveness of our algorithm both quantitatively and qualitatively on diverse compression methods.

[Feature Importance Ranking for Deep Learning](#)

- Maksymilian Wojtas, Ke Chen
- abstract@[open-review](#): Feature importance ranking has become a powerful tool for explainable AI. However, its nature of combinatorial optimization poses a great challenge for deep learning. In this paper, we propose a novel dual-net architecture consisting of operator and selector for discovery of an optimal feature subset of a fixed size and ranking the importance of those features in the optimal subset simultaneously. During learning, the operator is trained for a supervised learning task via optimal feature subset candidates generated by the selector that learns predicting the learning performance of the operator working on different optimal subset candidates. We develop an alternate learning algorithm that trains two nets jointly and incorporates a stochastic local search procedure into learning to address the combinatorial optimization challenge. In deployment, the selector generates an optimal feature subset and ranks feature importance, while the operator makes predictions based on the optimal subset for test data. A thorough evaluation on synthetic, benchmark and real data sets suggests that our approach outperforms several state-of-the-art feature importance ranking and supervised feature selection methods. (Our source code is available: <https://github.com/maksym33/FeatureImportanceDL>)

[Causal Estimation with Functional Confounders](#)

- Aahlad Puli, Adler Perotte, Rajesh Ranganath
- abstract@[open-review](#): Causal inference relies on two fundamental assumptions: ignorability and positivity. We study causal inference when the true confounder value can be expressed as a function of the observed data; we call this setting estimation with functional confounders (EFC). In this setting ignorability is satisfied, however positivity is violated, and causal inference is impossible in general. We consider two scenarios where causal effects are estimable. First, we discuss interventions on a part of the treatment called functional interventions and a sufficient condition for effect estimation of these interventions called functional positivity. Second, we develop conditions for nonparametric effect estimation based on the gradient fields of the functional confounder and the true outcome function. To estimate effects under these conditions, we develop Level-set Orthogonal Descent Estimation (LODE). Further, we prove error bounds on LODE's effect estimates, evaluate our methods on simulated and real data, and empirically demonstrate the value of EFC.

[Model Inversion Networks for Model-Based Optimization](#)

- Aviral Kumar, Sergey Levine
- abstract@[open-review](#): This work addresses data-driven optimization problems, where the goal is to find an input that maximizes an unknown score or reward function given access to a dataset of inputs with corresponding scores. When the inputs are high-dimensional and valid inputs constitute a small subset of this space (e.g., valid protein sequences or valid natural images), such model-based optimization problems become exceptionally difficult, since the optimizer must avoid out-of-distribution and invalid inputs. We propose to address such problems with model inversion networks (MINs), which learn an inverse mapping from scores to inputs. MINs can scale to high-dimensional input spaces and leverage offline logged data for both contextual and non-contextual optimization problems. MINs can also handle both purely offline data sources and active data collection. We evaluate MINs on high-dimensional model-based optimization problems over images, protein designs, and neural network controller parameters, and bandit optimization from logged data.

[Hausdorff Dimension, Heavy Tails, and Generalization in Neural Networks](#)

- Umut Simsekli, Ozan Sener, George Deligiannidis, Murat A. Erdogdu
- abstract@[open-review](#): Despite its success in a wide range of applications, characterizing the generalization properties of stochastic gradient descent (SGD) in non-convex deep learning problems is still an important challenge. While modeling the trajectories of SGD via stochastic differential equations (SDE) under heavy-tailed gradient noise has recently shed light over several peculiar characteristics of SGD, a rigorous treatment of the generalization properties of such SDEs in a learning theoretical framework is still missing. Aiming to bridge this gap, in this paper, we prove generalization bounds for SGD under the assumption that its trajectories can be well-approximated by a Feller process, which defines a rich class of Markov processes that include several recent SDE representations (both Brownian or heavy-tailed) as its special case. We show that the generalization error can be controlled by the Hausdorff dimension of the trajectories, which is intimately linked to the tail behavior of the driving process. Our results imply that heavier-

tailed processes should achieve better generalization; hence, the tail-index of the process can be used as a notion of ``capacity metric''. We support our theory with experiments on deep neural networks illustrating that the proposed capacity metric accurately estimates the generalization error, and it does not necessarily grow with the number of parameters unlike the existing capacity metrics in the literature.

[Exact expressions for double descent and implicit regularization via surrogate random design](#)

- Michal Derezhinski, Feynman T. Liang, Michael W. Mahoney
- abstract@[open-review](#): Double descent refers to the phase transition that is exhibited by the generalization error of unregularized learning models when varying the ratio between the number of parameters and the number of training samples. The recent success of highly over-parameterized machine learning models such as deep neural networks has motivated a theoretical analysis of the double descent phenomenon in classical models such as linear regression which can also generalize well in the over-parameterized regime. We provide the first exact non-asymptotic expressions for double descent of the minimum norm linear estimator. Our approach involves constructing a special determinantal point process which we call surrogate random design, to replace the standard i.i.d. design of the training sample. This surrogate design admits exact expressions for the mean squared error of the estimator while preserving the key properties of the standard design. We also establish an exact implicit regularization result for over-parameterized training samples. In particular, we show that, for the surrogate design, the implicit bias of the unregularized minimum norm estimator precisely corresponds to solving a ridge-regularized least squares problem on the population distribution. In our analysis we introduce a new mathematical tool of independent interest: the class of random matrices for which determinant commutes with expectation.

[Certifying Confidence via Randomized Smoothing](#)

- Aounon Kumar, Alexander Levine, Soheil Feizi, Tom Goldstein
- abstract@[open-review](#): Randomized smoothing has been shown to provide good certified-robustness guarantees for high-dimensional classification problems. It uses the probabilities of predicting the top two most-likely classes around an input point under a smoothing distribution to generate a certified radius for a classifier's prediction. However, most smoothing methods do not give us any information about the \emph{confidence} with which the underlying classifier (e.g., deep neural network) makes a prediction. In this work, we propose a method to generate certified radii for the prediction confidence of the smoothed classifier. We consider two notions for quantifying confidence: average prediction score of a class and the margin by which the average prediction score of one class exceeds that of another. We modify the Neyman-Pearson lemma (a key theorem in randomized smoothing) to design a procedure for computing the certified radius where the confidence is guaranteed to stay above a certain threshold. Our experimental results on CIFAR-10 and ImageNet datasets show that using information about the distribution of the confidence scores allows us to achieve a significantly better certified radius than ignoring it. Thus, we demonstrate that extra information about the base classifier at the input point can help improve certified guarantees for the smoothed classifier. Code for the experiments is available at \url{https://github.com/aounon/cdf-smoothing}.

[Learning Physical Constraints with Neural Projections](#)

- Shuqi Yang, Xingzhe He, Bo Zhu
- abstract@[open-review](#): We propose a new family of neural networks to predict the behaviors of physical systems by learning their underpinning constraints. A neural projection operator lies at the heart of our approach, composed of a lightweight network with an embedded recursive architecture that interactively enforces learned underpinning constraints and predicts the various governed behaviors of different physical systems. Our neural projection operator is motivated by the position-based dynamics model that has been used widely in game and visual effects industries to unify the various fast physics simulators. Our method can automatically and effectively uncover a broad range of constraints from observation point data, such as length, angle, bending, collision, boundary effects, and their arbitrary combinations, without any connectivity priors. We provide a multi-group point representation in conjunction with a configurable network connection mechanism to incorporate prior inputs for processing complex physical systems. We demonstrated the efficacy of our approach by learning a set of challenging physical systems all in a unified and simple fashion including: rigid bodies with complex geometries, ropes with varying length and bending, articulated soft and rigid bodies, and multi-object collisions with complex boundaries.

[Robust Optimization for Fairness with Noisy Protected Groups](#)

- Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, Michael Jordan
- abstract@[open-review](#): Many existing fairness criteria for machine learning involve equalizing some metric across protected groups such as race or gender. However, practitioners trying to audit or enforce such group-based criteria can easily face the problem of noisy or biased protected group information. First, we study the consequences of naively relying on noisy protected group labels: we provide an upper bound on the fairness violations on the true groups \$G\$ when the fairness criteria are satisfied on noisy groups \$\hat{G}\$. Second, we introduce two new approaches using robust optimization that, unlike the naive approach of only relying on \$\hat{G}\$, are guaranteed to satisfy fairness criteria on the true protected groups \$G\$ while minimizing a training objective. We provide theoretical guarantees that one such approach converges to an optimal feasible solution. Using two case studies, we show empirically that the robust approaches achieve better true group fairness guarantees than the naive approach.

[Noise-Contrastive Estimation for Multivariate Point Processes](#)

- Hongyuan Mei, Tom Wan, Jason Eisner
- abstract@[open-review](#): The log-likelihood of a generative model often involves both positive and negative terms. For a temporal multivariate point process, the negative term sums over all the possible event types at each time and also integrates over all the possible times. As a result, maximum likelihood estimation is expensive. We show how to instead apply a version of noise-contrastive estimation---a general parameter estimation method with a less expensive stochastic objective. Our specific instantiation of this general idea works out in an interestingly non-trivial way and has provable guarantees for its optimality, consistency and efficiency. On several synthetic and real-world datasets, our method shows benefits: for the model to achieve the same level of log-likelihood on held-out data, our method needs considerably fewer function evaluations and less wall-clock time.

[A Game-Theoretic Analysis of the Empirical Revenue Maximization Algorithm with Endogenous Sampling](#)

- Xiaotie Deng, Ron Lavi, Tao Lin, Qi Qi, Wenwei WANG, Xiang Yan
- abstract@[open-review](#): The Empirical Revenue Maximization (ERM) is one of the most important price learning algorithms in auction design: as the literature shows it can learn approximately optimal reserve prices for revenue-maximizing auctioneers in both repeated auctions and uniform-price auctions. However, in these applications the agents who provide inputs to ERM have incentives to manipulate the inputs to lower the outputted price. We generalize the definition of an incentive-awareness measure proposed by Lavi et al (2019), to quantify the reduction of ERM's outputted price due to a change of $m \geq 1$ out of N input samples, and provide specific convergence rates of this measure to zero as N goes to infinity for different types of input distributions. By adopting this measure, we construct an efficient, approximately incentive-compatible, and revenue-optimal learning algorithm using ERM in repeated auctions against non-myopic bidders, and show approximate group incentive-compatibility in uniform-price auctions.

[Neural Path Features and Neural Path Kernel : Understanding the role of gates in deep learning](#)

- Chandrashekhar Lakshminarayanan, Amit Vikram Singh

- abstract@[open-review](#): Rectified linear unit (ReLU) activations can also be thought of as 'gates', which, either pass or stop their pre-activation input when they are 'on' (when the pre-activation input is positive) or 'off' (when the pre-activation input is negative) respectively. A deep neural network (DNN) with ReLU activations has many gates, and the on/off status of each gate changes across input examples as well as network weights. For a given input example, only a subset of gates are 'active', i.e., on, and the sub-network of weights connected to these active gates is responsible for producing the output. At randomised initialisation, the active sub-network corresponding to a given input example is random. During training, as the weights are learnt, the active sub-networks are also learnt, and could hold valuable information. In this paper, we analytically characterise the role of gates and active sub-networks in deep learning. To this end, we encode the on/off state of the gates for a given input in a novel 'neural path feature' (NPF), and the weights of the DNN are encoded in a novel 'neural path value' (NPV). Further, we show that the output of network is indeed the inner product of NPF and NPV. The main result of the paper shows that the 'neural path kernel' associated with the NPF is a fundamental quantity that characterises the information stored in the gates of a DNN. We show via experiments (on MNIST and CIFAR-10) that in standard DNNs with ReLU activations NPFs are learnt during training and such learning is key for generalisation. Furthermore, NPFs and NPVs can be learnt in two separate networks and such learning also generalises well in experiments. In our experiments, we observe that almost all the information learnt by a DNN with ReLU activations is stored in the gates - a novel observation that underscores the need to investigate the role of the gates in DNNs.

[Multiscale Deep Equilibrium Models](#)

- Shaojie Bai, Vladlen Koltun, J. Zico Kolter
- abstract@[open-review](#): We propose a new class of implicit networks, the multiscale deep equilibrium model (MDEQ), suited to large-scale and highly hierarchical pattern recognition domains. An MDEQ directly solves for and backpropagates through the equilibrium points of multiple feature resolutions simultaneously, using implicit differentiation to avoid storing intermediate states (and thus requiring only $O(1)$ memory consumption). These simultaneously-learned multi-resolution features allow us to train a single model on a diverse set of tasks and loss functions, such as using a single MDEQ to perform both image classification and semantic segmentation. We illustrate the effectiveness of this approach on two large-scale vision tasks: ImageNet classification and semantic segmentation on high-resolution images from the Cityscapes dataset. In both settings, MDEQs are able to match or exceed the performance of recent competitive computer vision models: the first time such performance and scale have been achieved by an implicit deep learning approach. The code and pre-trained models are at <https://github.com/locuslab/mdeq>.

[Sparse Graphical Memory for Robust Planning](#)

- Scott Emmons, Ajay Jain, Misha Laskin, Thanard Kurutach, Pieter Abbeel, Deepak Pathak
- abstract@[open-review](#): To operate effectively in the real world, agents should be able to act from high-dimensional raw sensory input such as images and achieve diverse goals across long time-horizons. Current deep reinforcement and imitation learning methods can learn directly from high-dimensional inputs but do not scale well to long-horizon tasks. In contrast, classical graphical methods like A* search are able to solve long-horizon tasks, but assume that the state space is abstracted away from raw sensory input. Recent works have attempted to combine the strengths of deep learning and classical planning; however, dominant methods in this domain are still quite brittle and scale poorly with the size of the environment. We introduce Sparse Graphical Memory (SGM), a new data structure that stores states and feasible transitions in a sparse memory. SGM aggregates states according to a novel two-way consistency objective, adapting classic state aggregation criteria to goal-conditioned RL: two states are redundant when they are interchangeable both as goals and as starting states. Theoretically, we prove that merging nodes according to two-way consistency leads to an increase in shortest path lengths that scales only linearly with the merging threshold. Experimentally, we show that SGM significantly outperforms current state of the art methods on long horizon, sparse-reward visual navigation tasks. Project video and code are available at <https://sites.google.com/view/sparse-graphical-memory>.

[Second Order PAC-Bayesian Bounds for the Weighted Majority Vote](#)

- Andres Masegosa, Stephan Lorenzen, Christian Igel, Yevgeny Seldin
- abstract@[open-review](#): We present a novel analysis of the expected risk of weighted majority vote in multiclass classification. The analysis takes correlation of predictions by ensemble members into account and provides a bound that is amenable to efficient minimization, which yields improved weighting for the majority vote. We also provide a specialized version of our bound for binary classification, which allows to exploit additional unlabeled data for tighter risk estimation. In experiments, we apply the bound to improve weighting of trees in random forests and show that, in contrast to the commonly used first order bound, minimization of the new bound typically does not lead to degradation of the test error of the ensemble.

[Dirichlet Graph Variational Autoencoder](#)

- Jia Li, Jianwei Yu, Jiajin Li, Honglei Zhang, Kangfei Zhao, Yu Rong, Hong Cheng, Junzhou Huang
- abstract@[open-review](#): Graph Neural Networks (GNN) and Variational Autoencoders (VAEs) have been widely used in modeling and generating graphs with latent factors. However there is no clear explanation of what these latent factors are and why they perform well. In this work, we present Dirichlet Graph Variational Autoencoder (DGVAE) with graph cluster memberships as latent factors. Our study connects VAEs based graph generation and balanced graph cut, and provides a new way to understand and improve the internal mechanism of VAEs based graph generation. Specifically, we first interpret the reconstruction term of DGVAE as balanced graph cut in a principled way. Furthermore, motivated by the low pass characteristics in balanced graph cut, we propose a new variant of GNN named Heatts to encode the input graph into cluster memberships. Heatts utilizes the Taylor series for fast computation of Heat kernels and has better low pass characteristics than Graph Convolutional Networks (GCN). Through experiments on graph generation and graph clustering, we demonstrate the effectiveness of our proposed framework.

[Modeling Task Effects on Meaning Representation in the Brain via Zero-Shot MEG Prediction](#)

- Mariya Toneva, Otilia Stretcu, Barnabas Poczos, Leila Wehbe, Tom M. Mitchell
- abstract@[open-review](#): How meaning is represented in the brain is still one of the big open questions in neuroscience. Does a word (e.g., bird) always have the same representation, or does the task under which the word is processed alter its representation (answering "can you eat it?" versus "can it fly"?). The brain activity of subjects who read the same word while performing different semantic tasks has been shown to differ across tasks. However, it is still not understood how the task itself contributes to this difference. In the current work, we study Magnetoencephalography (MEG) brain recordings of participants tasked with answering questions about concrete nouns. We investigate the effect of the task (i.e. the question being asked) on the processing of the concrete noun by predicting the millisecond-resolution MEG recordings as a function of both the semantics of the noun and the task. Using this approach, we test several hypotheses about the task-stimulus interactions by comparing the zero-shot predictions made by these hypotheses for novel tasks and nouns not seen during training. We find that incorporating the task semantics significantly improves the prediction of MEG recordings, across participants. The improvement occurs 475–550ms after the participants first see the word, which corresponds to what is considered to be the ending time of semantic processing for a word. These results suggest that only the end of semantic processing of a word is task-dependent, and pose a challenge for future research to formulate new hypotheses for earlier task effects as a function of the task and stimuli.

[Counterfactual Vision-and-Language Navigation: Unravelling the Unseen](#)

- Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Javen Qinfeng Shi, Anton van den Hengel
- abstract@[open-review](#): The task of vision-and-language navigation (VLN) requires an agent to follow text instructions to find its way through simulated household environments. A prominent challenge is to train an agent capable of generalising to new environments at test time, rather than one that simply memorises trajectories and visual details observed during training. We propose a new learning strategy that learns both from observations and generated

counterfactual environments. We describe an effective algorithm to generate counterfactual observations on the fly for VLN, as linear combinations of existing environments. Simultaneously, we encourage the agent's actions to remain stable between original and counterfactual environments through our novel training objective-effectively removing the spurious features that otherwise bias the agent. Our experiments show that this technique provides significant improvements in generalisation on benchmarks for Room-to-Room navigation and Embodied Question Answering.

[Robust Quantization: One Model to Rule Them All](#)

- moran shkolnik, Brian Chmiel, Ron Banner, Gil Shomron, Yury Nahshan, Alex Bronstein, Uri Weiser
- abstract@[open-review](#): Neural network quantization methods often involve simulating the quantization process during training, making the trained model highly dependent on the target bit-width and precise way quantization is performed. Robust quantization offers an alternative approach with improved tolerance to different classes of data-types and quantization policies. It opens up new exciting applications where the quantization process is not static and can vary to meet different circumstances and implementations. To address this issue, we propose a method that provides intrinsic robustness to the model against a broad range of quantization processes. Our method is motivated by theoretical arguments and enables us to store a single generic model capable of operating at various bit-widths and quantization policies. We validate our method's effectiveness on different ImageNet Models. A reference implementation accompanies the paper.

[Enabling certification of verification-agnostic networks via memory-efficient semidefinite programming](#)

- Sumanth Dathathri, Krishnamurthy Dvijotham, Alexey Kurakin, Aditi Raghunathan, Jonathan Uesato, Rudy R. Bunel, Shreya Shankar, Jacob Steinhardt, Ian Goodfellow, Percy S. Liang, Pushmeet Kohli
- abstract@[open-review](#): Convex relaxations have emerged as a promising approach for verifying properties of neural networks, but widely used using Linear Programming (LP) relaxations only provide meaningful certificates when networks are specifically trained to facilitate verification. This precludes many important applications which involve \emph{verification-agnostic} networks that are not trained specifically to promote verifiability. On the other hand, semidefinite programming (SDP) relaxations have shown success on verification-agnostic networks, such as adversarially trained image classifiers without additional regularization, but do not currently scale beyond small networks due to poor time and space asymptotics. In this work, we propose a first-order dual SDP algorithm that provides (1) any-time bounds (2) requires memory only linear in the total number of network activations and (3) has per-iteration complexity that scales linearly with the complexity of a forward/backward pass through the network. By exploiting iterative eigenvector methods, we express all solver operations in terms of forward and backward passes through the network, enabling efficient use of hardware optimized for deep learning. This allows us to dramatically improve the magnitude of $\|\cdot\|_{\infty}$ perturbations for which we can verify robustness verification-agnostic networks (\$1\% \rightarrow 88\%\$ on MNIST, \$6\% \rightarrow 40\%\$ on CIFAR-10). We also demonstrate tight verification for a quadratic stability specification for the decoder of a variational autoencoder.

[Federated Accelerated Stochastic Gradient Descent](#)

- Honglin Yuan, Tengyu Ma
- abstract@[open-review](#): We propose Federated Accelerated Stochastic Gradient Descent (FedAc), a principled acceleration of Federated Averaging (FedAvg, also known as Local SGD) for distributed optimization. FedAc is the first provable acceleration of FedAvg that improves convergence speed and communication efficiency on various types of convex functions. For example, for strongly convex and smooth functions, when using M workers, the previous state-of-the-art FedAvg analysis can achieve a linear speedup in M if given M rounds of synchronization, whereas FedAc only requires $M^{1/3}$ rounds. Moreover, we prove stronger guarantees for FedAc when the objectives are third-order smooth. Our technique is based on a potential-based perturbed iterate analysis, a novel stability analysis of generalized accelerated SGD, and a strategic tradeoff between acceleration and stability.

[Robust Density Estimation under Besov IPM Losses](#)

- Ananya Uppal, Shashank Singh, Barnabas Poczos
- abstract@[open-review](#): We study minimax convergence rates of nonparametric density estimation under the Huber contamination model, in which a ``contaminated'' proportion of the data comes from an unknown outlier distribution. We provide the first results for this problem under a large family of losses, called Besov integral probability metrics (IPMs), that include L^p , Wasserstein, Kolmogorov-Smirnov, Cramer-von Mises, and other commonly used metrics. Under a range of smoothness assumptions on the population and outlier distributions, we show that a re-scaled thresholding wavelet estimator converges at the minimax optimal rate under a wide variety of losses and also exhibits optimal dependence on the contamination proportion. We also provide a purely data-dependent extension of the estimator that adapts to both an unknown contamination proportion and the unknown smoothness of the true density. Finally, based on connections shown recently between density estimation under IPM losses and generative adversarial networks (GANs), we show that certain GAN architectures are robustly minimax optimal.

[An analytic theory of shallow networks dynamics for hinge loss classification](#)

- Franco Pellegrini, Giulio Biroli
- abstract@[open-review](#): Neural networks have been shown to perform incredibly well in classification tasks over structured high-dimensional datasets. However, the learning dynamics of such networks is still poorly understood. In this paper we study in detail the training dynamics of a simple type of neural network: a single hidden layer trained to perform a classification task. We show that in a suitable mean-field limit this case maps to a single-node learning problem with a time-dependent dataset determined self-consistently from the average nodes population. We specialize our theory to the prototypical case of a linearly separable dataset and a linear hinge loss, for which the dynamics can be explicitly solved in the infinite dataset limit. This allows us to address in a simple setting several phenomena appearing in modern networks such as slowing down of training dynamics, crossover between feature and lazy learning, and overfitting. Finally, we assess the limitations of mean-field theory by studying the case of large but finite number of nodes and of training samples.

[Fixed-Support Wasserstein Barycenters: Computational Hardness and Fast Algorithm](#)

- Tianyi Lin, Nhat Ho, Xi Chen, Marco Cuturi, Michael Jordan
- abstract@[open-review](#): We study the fixed-support Wasserstein barycenter problem (FS-WBP), which consists in computing the Wasserstein barycenter of m discrete probability measures supported on a finite metric space of size n . We show first that the constraint matrix arising from the standard linear programming (LP) representation of the FS-WBP is not totally unimodular when $m \geq 3$ and $n \geq 3$. This result resolves an open question pertaining to the relationship between the FS-WBP and the minimum-cost flow (MCF) problem since it proves that the FS-WBP in the standard LP form is not an MCF problem when $m \geq 3$ and $n \geq 3$. We also develop a provably fast deterministic variant of the celebrated iterative Bregman projection (IBP) algorithm, named \textsc{FastIBP}, with a complexity bound of $\tilde{O}(mn^{7/3}\varepsilon^{-4/3})$, where $\varepsilon \in (0, 1)$ is the desired tolerance. This complexity bound is better than the best known complexity bound of $\tilde{O}(mn^2\varepsilon^{-2})$ for the IBP algorithm in terms of ε , and that of $\tilde{O}(mn^{5/2}\varepsilon^{-1})$ from accelerated alternating minimization algorithm or accelerated primal-dual adaptive gradient algorithm in terms of n . Finally, we conduct extensive experiments with both synthetic data and real images and demonstrate the favorable performance of the \textsc{FastIBP} algorithm in practice.

[Learning to Orient Surfaces by Self-supervised Spherical CNNs](#)

- Riccardo Spezialetti, Federico Stella, Marlon Marcon, Luciano Silva, Samuele Salti, Luigi Di Stefano
- abstract@[open-review](#): Defining and reliably finding a canonical orientation for 3D surfaces is key to many Computer Vision and Robotics applications. This task is commonly addressed by handcrafted algorithms exploiting geometric cues deemed as distinctive and robust by the designer. Yet, one might conjecture that humans learn the notion of the inherent orientation of 3D objects from experience and that machines may do so alike. In this work, we show the feasibility of learning a robust canonical orientation for surfaces represented as point clouds. Based on the observation that the quintessential property of a canonical orientation is equivariance to 3D rotations, we propose to employ Spherical CNNs, a recently introduced machinery that can learn equivariant representations defined on the Special Orthogonal group $SO(3)$. Specifically, spherical correlations compute feature maps whose elements define 3D rotations. Our method learns such feature maps from raw data by a self-supervised training procedure and robustly selects a rotation to transform the input point cloud into a learned canonical orientation. Thereby, we realize the first end-to-end learning approach to define and extract the canonical orientation of 3D shapes, which we aptly dub Compass. Experiments on several public datasets prove its effectiveness at orienting local surface patches as well as whole objects.

[Adam with Bandit Sampling for Deep Learning](#)

- Rui Liu, Tianyi Wu, Barzan Mozafari
- abstract@[open-review](#): Adam is a widely used optimization method for training deep learning models. It computes individual adaptive learning rates for different parameters. In this paper, we propose a generalization of Adam, called Adambs, that allows us to also adapt to different training examples based on their importance in the model's convergence. To achieve this, we maintain a distribution over all examples, selecting a mini-batch in each iteration by sampling according to this distribution, which we update using a multi-armed bandit algorithm. This ensures that examples that are more beneficial to the model training are sampled with higher probabilities. We theoretically show that Adambs improves the convergence rate of Adam--- $\$O(\sqrt{\frac{\log n}{T}})$ instead of $\$O(\sqrt{\frac{n}{T}})$ in some cases. Experiments on various models and datasets demonstrate Adambs's fast convergence in practice.

[Parabolic Approximation Line Search for DNNs](#)

- Maximus Mutschler, Andreas Zell
- abstract@[open-review](#): A major challenge in current optimization research for deep learning is to automatically find optimal step sizes for each update step. The optimal step size is closely related to the shape of the loss in the update step direction. However, this shape has not yet been examined in detail. This work shows empirically that the sample loss over lines in negative gradient direction is mostly convex and well suited for one-dimensional parabolic approximations. Exploiting this parabolic property we introduce a simple and robust line search approach, which performs loss-shape dependent update steps. Our approach combines well-known methods such as parabolic approximation, line search and conjugate gradient, to perform efficiently. It successfully competes with common and state-of-the-art optimization methods on a large variety of experiments without the need of hand-designed step size schedules. Thus, it is of interest for objectives where step-size schedules are unknown or do not perform well. Our excessive evaluation includes multiple comprehensive hyperparameter grid searches on several datasets and architectures. We provide proof of convergence for an adapted scenario. Finally, we give a general investigation of exact line searches in the context of sample losses and exact losses, including their relation to our line search approach.

[Agnostic Learning of a Single Neuron with Gradient Descent](#)

- Spencer Frei, Yuan Cao, Quanquan Gu
- abstract@[open-review](#): We consider the problem of learning the best-fitting single neuron as measured by the expected square loss $\mathbb{E}_{\{(x,y)\sim \mathcal{D}\}}[(\sigma(w^\top x) - y)^2]$ over some unknown joint distribution \mathcal{D} by using gradient descent to minimize the empirical risk induced by a set of i.i.d. samples $S \sim \mathcal{D}^n$. The activation function σ is an arbitrary Lipschitz and non-decreasing function, making the optimization problem nonconvex and nonsmooth in general, and covers typical neural network activation functions and inverse link functions in the generalized linear model setting. In the agnostic PAC learning setting, where no assumption on the relationship between the labels y and the input x is made, if the optimal population risk is OPT , we show that gradient descent achieves population risk $\mathcal{O}(\mathsf{OPT}) + \epsilon$ in polynomial time and sample complexity when σ is strictly increasing. For the ReLU activation, our population risk guarantee is $\mathcal{O}(\mathsf{OPT}^{1/2}) + \epsilon$. When labels take the form $y = \sigma(v^\top x) + \xi$ for zero-mean sub-Gaussian noise ξ , we show that the population risk guarantees for gradient descent improve to $\mathsf{OPT} + \epsilon$. Our sample complexity and runtime guarantees are (almost) dimension independent, and when σ is strictly increasing, require no distributional assumptions beyond boundedness. For ReLU, we show the same results under a nondegeneracy assumption for the marginal distribution of the input.

[Statistical Efficiency of Thompson Sampling for Combinatorial Semi-Bandits](#)

- Pierre Perrault, Etienne Boursier, Michal Valko, Vianney Perchet
- abstract@[open-review](#): We investigate stochastic combinatorial multi-armed bandit with semi-bandit feedback (CMAB). In CMAB, the question of the existence of an efficient policy with an optimal asymptotic regret (up to a factor poly-logarithmic with the action size) is still open for many families of distributions, including mutually independent outcomes, and more generally the multivariate sub-Gaussian family. We propose to answer the above question for these two families by analyzing variants of the Combinatorial Thompson Sampling policy (CTS). For mutually independent outcomes in $[0,1]$, we propose a tight analysis of CTS using Beta priors. We then look at the more general setting of multivariate sub-Gaussian outcomes and propose a tight analysis of CTS using Gaussian priors. This last result gives us an alternative to the Efficient Sampling for Combinatorial Bandit policy (ESCB), which, although optimal, is not computationally efficient.

[Analytic Characterization of the Hessian in Shallow ReLU Models: A Tale of Symmetry](#)

- Yossi Arjevani, Michael Field
- abstract@[open-review](#): We consider the optimization problem associated with fitting two-layers ReLU networks with respect to the squared loss, where labels are generated by a target network. We leverage the rich symmetry structure to analytically characterize the Hessian at various families of spurious minima in the natural regime where the number of inputs d and the number of hidden neurons k is finite. In particular, we prove that for $d \geq k$ standard Gaussian inputs: (a) of the dk eigenvalues of the Hessian, $dk - O(d)$ concentrate near zero, (b) $\Omega(d)$ of the eigenvalues grow linearly with k . Although this phenomenon of extremely skewed spectrum has been observed many times before, to our knowledge, this is the first time it has been established rigorously. Our analytic approach uses techniques, new to the field, from symmetry breaking and representation theory, and carries important implications for our ability to argue about statistical generalization through local curvature.

[Generative causal explanations of black-box classifiers](#)

- Matthew O'Shaughnessy, Gregory Canal, Marissa Connor, Christopher Rozell, Mark Davenport
- abstract@[open-review](#): We develop a method for generating causal post-hoc explanations of black-box classifiers based on a learned low-dimensional representation of the data. The explanation is causal in the sense that changing learned latent factors produces a change in the classifier output statistics. To construct these explanations, we design a learning framework that leverages a generative model and information-theoretic measures of causal influence. Our objective function encourages both the generative model to faithfully represent the data distribution and the latent factors to have a large causal influence on the classifier output. Our method learns both global and local explanations, is compatible with any classifier that admits class probabilities

and a gradient, and does not require labeled attributes or knowledge of causal structure. Using carefully controlled test cases, we provide intuition that illuminates the function of our causal objective. We then demonstrate the practical utility of our method on image recognition tasks.

[Sub-sampling for Efficient Non-Parametric Bandit Exploration](#)

- Dorian Baudry, Emilie Kaufmann, Odalric-Ambrym Maillard
- abstract@[open-review](#): In this paper we propose the first multi-armed bandit algorithm based on re-sampling that achieves asymptotically optimal regret simultaneously for different families of arms (namely Bernoulli, Gaussian and Poisson distributions). Unlike Thompson Sampling which requires to specify a different prior to be optimal in each case, our proposal RB-SDA does not need any distribution-dependent tuning. RB-SDA belongs to the family of Sub-sampling Duelling Algorithms (SDA) which combines the sub-sampling idea first used by the BESA and SSMC algorithms with different sub-sampling schemes. In particular, RB-SDA uses Random Block sampling. We perform an experimental study assessing the flexibility and robustness of this promising novel approach for exploration in bandit models.

[Learning under Model Misspecification: Applications to Variational and Ensemble methods](#)

- Andres Masegosa
- abstract@[open-review](#): Virtually any model we use in machine learning to make predictions does not perfectly represent reality. So, most of the learning happens under model misspecification. In this work, we present a novel analysis of the generalization performance of Bayesian model averaging under model misspecification and i.i.d. data using a new family of second-order PAC-Bayes bounds. This analysis shows, in simple and intuitive terms, that Bayesian model averaging provides suboptimal generalization performance when the model is misspecified. In consequence, we provide strong theoretical arguments showing that Bayesian methods are not optimal for learning predictive models, unless the model class is perfectly specified. Using novel second-order PAC-Bayes bounds, we derive a new family of Bayesian-like algorithms, which can be implemented as variational and ensemble methods. The output of these algorithms is a new posterior distribution, different from the Bayesian posterior, which induces a posterior predictive distribution with better generalization performance. Experiments with Bayesian neural networks illustrate these findings.

[Language Through a Prism: A Spectral Approach for Multiscale Language Representations](#)

- Alex Tamkin, Dan Jurafsky, Noah Goodman
- abstract@[open-review](#): Language exhibits structure at a wide range of scales, from subwords to words, sentences, paragraphs, and documents. We propose building models that isolate scale-specific information in deep representations, and develop methods for encouraging models during training to learn more about particular scales of interest. Our method for creating scale-specific neurons in deep NLP models constrains how the activation of a neuron can change across the tokens of an input by interpreting those activations as a digital signal and filtering out parts of its frequency spectrum. This technique enables us to extract scale-specific information from BERT representations: by filtering out different frequencies we can produce new representations that perform well on part of speech tagging (word-level), dialog speech acts classification (utterance-level), or topic classification (document-level), while performing poorly on the other tasks. We also present a prism layer for use during training, which constrains different neurons of a BERT model to different parts of the frequency spectrum. Our proposed BERT + Prism model is better able to predict masked tokens using long-range context, and produces individual multiscale representations that perform with comparable or improved performance across all three tasks. Our methods are general and readily applicable to other domains besides language, such as images, audio, and video.

[DVERGE: Diversifying Vulnerabilities for Enhanced Robust Generation of Ensembles](#)

- Huanrui Yang, Jingyang Zhang, Hongliang Dong, Nathan Inkawich, Andrew Gardner, Andrew Touchet, Wesley Wilkes, Heath Berry, Hai Li
- abstract@[open-review](#): Recent research finds CNN models for image classification demonstrate overlapped adversarial vulnerabilities: adversarial attacks can mislead CNN models with small perturbations, which can effectively transfer between different models trained on the same dataset. Adversarial training, as a general robustness improvement technique, eliminates the vulnerability in a single model by forcing it to learn robust features. The process is hard, often requires models with large capacity, and suffers from significant loss on clean data accuracy. Alternatively, ensemble methods are proposed to induce sub-models with diverse outputs against a transfer adversarial example, making the ensemble robust against transfer attacks even if each sub-model is individually non-robust. Only small clean accuracy drop is observed in the process. However, previous ensemble training methods are not efficacious in inducing such diversity and thus ineffective on reaching robust ensemble. We propose DVERGE, which isolates the adversarial vulnerability in each sub-model by distilling non-robust features, and diversifies the adversarial vulnerability to induce diverse outputs against a transfer attack. The novel diversity metric and training procedure enables DVERGE to achieve higher robustness against transfer attacks comparing to previous ensemble methods, and enables the improved robustness when more sub-models are added to the ensemble. The code of this work is available at <https://github.com/zjysteven/DVERGE>.

[Towards practical differentially private causal graph discovery](#)

- Lun Wang, Qi Pang, Dawn Song
- abstract@[open-review](#): Causal graph discovery refers to the process of discovering causal relation graphs from purely observational data. Like other statistical data, a causal graph might leak sensitive information about participants in the dataset. In this paper, we present a differentially private causal graph discovery algorithm, Priv-PC, which improves both utility and running time compared to the state-of-the-art. The design of Priv-PC follows a novel paradigm called sieve-and-examine which uses a small amount of privacy budget to filter out “insignificant” queries, and leverages the remaining budget to obtain highly accurate answers for the “significant” queries. We also conducted the first sensitivity analysis for conditional independence tests including conditional Kendall’s τ , and conditional Spearman’s ρ . We evaluated Priv-PC on 7 public datasets and compared with the state-of-the-art. The results show that Priv-PC achieves 10.61 to 293.87 times speedup and better utility. The implementation of Priv-PC, including the code used in our evaluation, is available at <https://github.com/sunblaze-ucb/Priv-PC-Differentially-Private-Causal-Graph-Discovery>.

[Independent Policy Gradient Methods for Competitive Reinforcement Learning](#)

- Constantinos Daskalakis, Dylan J. Foster, Noah Golowich
- abstract@[open-review](#): We obtain global, non-asymptotic convergence guarantees for independent learning algorithms in competitive reinforcement learning settings with two agents (i.e., zero-sum stochastic games). We consider an episodic setting where in each episode, each player independently selects a policy and observes only their own actions and rewards, along with the state. We show that if both players run policy gradient methods in tandem, their policies will converge to a min-max equilibrium of the game, as long as their learning rates follow a two-timescale rule (which is necessary). To the best of our knowledge, this constitutes the first finite-sample convergence result for independent learning in competitive RL, as prior work has largely focused on centralized/coordinated procedures for equilibrium computation.

[The Value Equivalence Principle for Model-Based Reinforcement Learning](#)

- Christopher Grimm, Andre Barreto, Satinder Singh, David Silver
- abstract@[open-review](#): Learning models of the environment from data is often viewed as an essential component to building intelligent reinforcement learning (RL) agents. The common practice is to separate the learning of the model from its use, by constructing a model of the environment’s

dynamics that correctly predicts the observed state transitions. In this paper we argue that the limited representational resources of model-based RL agents are better used to build models that are directly useful for value-based planning. As our main contribution, we introduce the principle of value equivalence: two models are value equivalent with respect to a set of functions and policies if they yield the same Bellman updates. We propose a formulation of the model learning problem based on the value equivalence principle and analyze how the set of feasible solutions is impacted by the choice of policies and functions. Specifically, we show that, as we augment the set of policies and functions considered, the class of value equivalent models shrinks, until eventually collapsing to a single point corresponding to a model that perfectly describes the environment. In many problems, directly modelling state-to-state transitions may be both difficult and unnecessary. By leveraging the value-equivalence principle one may find simpler models without compromising performance, saving computation and memory. We illustrate the benefits of value-equivalent model learning with experiments comparing it against more traditional counterparts like maximum likelihood estimation. More generally, we argue that the principle of value equivalence underlies a number of recent empirical successes in RL, such as Value Iteration Networks, the Predictron, Value Prediction Networks, TreeQN, and MuZero, and provides a first theoretical underpinning of those results.

[Structured Convolutions for Efficient Neural Network Design](#)

- Yash Bhalgat, Yizhe Zhang, Jamie Menjay Lin, Fatih Porikli
- abstract@[open-review](#): In this work, we tackle model efficiency by exploiting redundancy in the implicit structure of the building blocks of convolutional neural networks. We start our analysis by introducing a general definition of Composite Kernel structures that enable the execution of convolution operations in the form of efficient, scaled, sum-pooling components. As its special case, we propose Structured Convolutions and show that these allow decomposition of the convolution operation into a sum-pooling operation followed by a convolution with significantly lower complexity and fewer weights. We show how this decomposition can be applied to 2D and 3D kernels as well as the fully-connected layers. Furthermore, we present a Structural Regularization loss that promotes neural network layers to leverage on this desired structure in a way that, after training, they can be decomposed with negligible performance loss. By applying our method to a wide range of CNN architectures, we demonstrate 'structured' versions of the ResNets that are up to 2x smaller and a new Structured-MobileNetV2 that is more efficient while staying within an accuracy loss of 1% on ImageNet and CIFAR-10 datasets. We also show similar structured versions of EfficientNet on ImageNet and HRNet architecture for semantic segmentation on the Cityscapes dataset. Our method performs equally well or superior in terms of the complexity reduction in comparison to the existing tensor decomposition and channel pruning methods.

[Latent World Models For Intrinsically Motivated Exploration](#)

- Aleksandr Ermolov, Nicu Sebe
- abstract@[open-review](#): In this work we consider partially observable environments with sparse rewards. We present a self-supervised representation learning method for image-based observations, which arranges embeddings respecting temporal distance of observations. This representation is empirically robust to stochasticity and suitable for novelty detection from the error of a predictive forward model. We consider episodic and life-long uncertainties to guide the exploration. We propose to estimate the missing information about the environment with the world model, which operates in the learned latent space. As a motivation of the method, we analyse the exploration problem in a tabular Partially Observable Labyrinth. We demonstrate the method on image-based hard exploration environments from the Atari benchmark and report significant improvement with respect to prior work. The source code of the method and all the experiments is available at <https://github.com/htdt/lwm>.

[Estimating Rank-One Spikes from Heavy-Tailed Noise via Self-Avoiding Walks](#)

- Jingqiu Ding, Samuel Hopkins, David Steurer
- abstract@[open-review](#): We study symmetric spiked matrix models with respect to a general class of noise distributions. Given a rank-1 deformation of a random noise matrix, whose entries are independently distributed with zero mean and unit variance, the goal is to estimate the rank-1 part. For the case of Gaussian noise, the top eigenvector of the given matrix is a widely-studied estimator known to achieve optimal statistical guarantees, e.g., in the sense of the celebrated BBP phase transition. However, this estimator can fail completely for heavy-tailed noise. In this work, we exhibit an estimator that works for heavy-tailed noise up to the BBP threshold that is optimal even for Gaussian noise. We give a non-asymptotic analysis of our estimator which relies only on the variance of each entry remaining constant as the size of the matrix grows: higher moments may grow arbitrarily fast or even fail to exist. Previously, it was only known how to achieve these guarantees if higher-order moments of the noises are bounded by a constant independent of the size of the matrix. Our estimator can be evaluated in polynomial time by counting self-avoiding walks via a color coding technique. Moreover, we extend our estimator to spiked tensor models and establish analogous results.

[Policy Improvement via Imitation of Multiple Oracles](#)

- Ching-An Cheng, Andrey Kolobov, Alekh Agarwal
- abstract@[open-review](#): Despite its promise, reinforcement learning's real-world adoption has been hampered by the need for costly exploration to learn a good policy. Imitation learning (IL) mitigates this shortcoming by using an oracle policy during training as a bootstrap to accelerate the learning process. However, in many practical situations, the learner has access to multiple suboptimal oracles, which may provide conflicting advice in a state. The existing IL literature provides a limited treatment of such scenarios. Whereas in the single-oracle case, the return of the oracle's policy provides an obvious benchmark for the learner to compete against, neither such a benchmark nor principled ways of outperforming it are known for the multi-oracle setting. In this paper, we propose the state-wise maximum of the oracle policies' values as a natural baseline to resolve conflicting advice from multiple oracles. Using a reduction of policy optimization to online learning, we introduce a novel IL algorithm MAMBA, which can provably learn a policy competitive with this benchmark. In particular, MAMBA optimizes policies by using a gradient estimator in the style of generalized advantage estimation (GAE). Our theoretical analysis shows that this design makes MAMBA robust and enables it to outperform the oracle policies by a larger margin than the IL state of the art, even in the single-oracle case. In an evaluation against standard policy gradient with GAE and AggreVaTe(D), we showcase MAMBA's ability to leverage demonstrations both from a single and from multiple weak oracles, and significantly speed up policy optimization.

[Training Generative Adversarial Networks by Solving Ordinary Differential Equations](#)

- Chongli Qin, Yan Wu, Jost Tobias Springenberg, Andy Brock, Jeff Donahue, Timothy Lillicrap, Pushmeet Kohli
- abstract@[open-review](#): The instability of Generative Adversarial Network (GAN) training has frequently been attributed to gradient descent. Consequently, recent methods have aimed to tailor the models and training procedures to stabilise the discrete updates. In contrast, we study the continuous-time dynamics induced by GAN training. Both theory and toy experiments suggest that these dynamics are in fact surprisingly stable. From this perspective, we hypothesise that instabilities in training GANs arise from the integration error in discretising the continuous dynamics. We experimentally verify that well-known ODE solvers (such as Runge-Kutta) can stabilise training - when combined with a regulariser that controls the integration error. Our approach represents a radical departure from previous methods which typically use adaptive optimisation and stabilisation techniques that constrain the functional space (e.g. Spectral Normalisation). Evaluation on CIFAR-10 and ImageNet shows that our method outperforms several strong baselines, demonstrating its efficacy.

[Learning of Discrete Graphical Models with Neural Networks](#)

- Abhijith Jayakumar, Andrey Lokhov, Sidhant Misra, Marc Vuffray

- abstract@[open-review](#): Graphical models are widely used in science to represent joint probability distributions with an underlying conditional dependence structure. The inverse problem of learning a discrete graphical model given i.i.d samples from its joint distribution can be solved with near-optimal sample complexity using a convex optimization method known as Generalized Regularized Interaction Screening Estimator (GRISE). But the computational cost of GRISE becomes prohibitive when the energy function of the true graphical model has higher order terms. We introduce NeurISE, a neural net based algorithm for graphical model learning, to tackle this limitation of GRISE. We use neural nets as function approximators in an Interaction Screening objective function. The optimization of this objective then produces a neural-net representation for the conditionals of the graphical model. NeurISE algorithm is seen to be a better alternative to GRISE when the energy function of the true model has a high order with a high degree of symmetry. In these cases NeurISE is able to find the correct parsimonious representation for the conditionals without being fed any prior information about the true model. NeurISE can also be used to learn the underlying structure of the true model with some simple modifications to its training procedure. In addition, we also show a variant of NeurISE that can be used to learn a neural net representation for the full energy function of the true model.

[RepPoints v2: Verification Meets Regression for Object Detection](#)

- Yihong Chen, Zheng Zhang, Yue Cao, Liwei Wang, Stephen Lin, Han Hu
- abstract@[open-review](#): Verification and regression are two general methodologies for prediction in neural networks. Each has its own strengths: verification can be easier to infer accurately, and regression is more efficient and applicable to continuous target variables. Hence, it is often beneficial to carefully combine them to take advantage of their benefits. In this paper, we take this philosophy to improve state-of-the-art object detection, specifically by RepPoints. Though RepPoints provides high performance, we find that its heavy reliance on regression for object localization leaves room for improvement. We introduce verification tasks into the localization prediction of RepPoints, producing RepPoints v2, which proves consistent improvements of about 2.0 mAP over the original RepPoints on COCO object detection benchmark using different backbones and training methods. RepPoints v2 also achieves 52.1 mAP on the COCO \texttt{test-dev} by a single model. Moreover, we show that the proposed approach can more generally elevate other object detection frameworks as well as applications such as instance segmentation.

[Unfolding the Alternating Optimization for Blind Super Resolution](#)

- zhengxiong luo, Yan Huang, Shang Li, Liang Wang, Tieniu Tan
- abstract@[open-review](#): Previous methods decompose blind super resolution (SR) problem into two sequential steps: \textit{i}) estimating blur kernel from given low-resolution (LR) image and \textit{ii}) restoring SR image based on estimated kernel. This two-step solution involves two independently trained models, which may not well compatible with each other. Small estimation error of the first step could cause severe performance drop of the second one. While on the other hand, the first step can only utilize limited information from LR image, which makes it difficult to predict highly accurate blur kernel. Towards these issues, instead of considering these two steps separately, we adopt an alternating optimization algorithm, which can estimate blur kernel and restore SR image in a single model. Specifically, we design two convolutional neural modules, namely \textit{Restorer} and \textit{Estimator}. \textit{Restorer} restores SR image based on predicted kernel, and \textit{Estimator} estimates blur kernel with the help of restored SR image. We alternate these two modules repeatedly and unfold this process to form an end-to-end trainable network. In this way, \textit{Estimator} utilizes information from both LR and SR images, which makes the estimation of blur kernel easier. More importantly, \textit{Restorer} is trained with the kernel estimated by \textit{Estimator}, instead of ground-truth kernel, thus \textit{Restorer} could be more tolerant to the estimation error of \textit{Estimator}. Extensive experiments on synthetic datasets and real-world images show that our model can largely outperform state-of-the-art methods and produce more visually favorable results at much higher speed. The source code will be publicly available.

[Entrywise convergence of iterative methods for eigenproblems](#)

- Vasileios Charisopoulos, Austin R. Benson, Anil Damle
- abstract@[open-review](#): Several problems in machine learning, statistics, and other fields rely on computing eigenvectors. For large scale problems, the computation of these eigenvectors is typically performed via iterative schemes such as subspace iteration or Krylov methods. While there is classical and comprehensive analysis for subspace convergence guarantees with respect to the spectral norm, in many modern applications other notions of subspace distance are more appropriate. Recent theoretical work has focused on perturbations of subspaces measured in the $\|\cdot\|_2$ norm, but does not consider the actual computation of eigenvectors. Here we address the convergence of subspace iteration when distances are measured in the $\|\cdot\|_2$ norm and provide deterministic bounds. We complement our analysis with a practical stopping criterion and demonstrate its applicability via numerical experiments. Our results show that one can get comparable performance on downstream tasks while requiring fewer iterations, thereby saving substantial computational time.

[Learning Object-Centric Representations of Multi-Object Scenes from Multiple Views](#)

- Nanbo Li, Cian Eastwood, Robert Fisher
- abstract@[open-review](#): Learning object-centric representations of multi-object scenes is a promising approach towards machine intelligence, facilitating high-level reasoning and control from visual sensory data. However, current approaches for \textit{unsupervised object-centric scene representation} are incapable of aggregating information from multiple observations of a scene. As a result, these ``single-view'' methods form their representations of a 3D scene based only on a single 2D observation (view). Naturally, this leads to several inaccuracies, with these methods falling victim to single-view spatial ambiguities. To address this, we propose \textit{The Multi-View and Multi-Object Network (MulMON)}---a method for learning accurate, object-centric representations of multi-object scenes by leveraging multiple views. In order to sidestep the main technical difficulty of the \textit{multi-object-multi-view} scenario---maintaining object correspondences across views---MulMON iteratively updates the latent object representations for a scene over multiple views. To ensure that these iterative updates do indeed aggregate spatial information to form a complete 3D scene understanding, MulMON is asked to predict the appearance of the scene from novel viewpoints during training. Through experiments we show that MulMON better-resolves spatial ambiguities than single-view methods---learning more accurate and disentangled object representations---and also achieves new functionality in predicting object segmentations for novel viewpoints.

[A Catalyst Framework for Minimax Optimization](#)

- Junchi Yang, Siqi Zhang, Negar Kiyavash, Niao He
- abstract@[open-review](#): We introduce a generic \textit{two-loop} scheme for smooth minimax optimization with strongly-convex-concave objectives. Our approach applies the accelerated proximal point framework (or Catalyst) to the associated \textit{dual problem} and takes full advantage of existing gradient-based algorithms to solve a sequence of well-balanced strongly-convex-strongly-concave minimax problems. Despite its simplicity, this leads to a family of near-optimal algorithms with improved complexity over all existing methods designed for strongly-convex-concave minimax problems. Additionally, we obtain the first variance-reduced algorithms for this class of minimax problems with finite-sum structure and establish even faster convergence rate. Furthermore, when extended to the nonconvex-concave minimax optimization, our algorithm again achieves the state-of-the-art complexity for finding a stationary point. We carry out several numerical experiments showcasing the superiority of the Catalyst framework in practice.

[Self-supervised Co-Training for Video Representation Learning](#)

- Tengda Han, Weidi Xie, Andrew Zisserman
- abstract@[open-review](#): The objective of this paper is visual-only self-supervised video representation learning. We make the following contributions: (i) we investigate the benefit of adding semantic-class positives to instance-based Info Noise Contrastive Estimation (InfoNCE) training, showing that this

form of supervised contrastive learning leads to a clear improvement in performance; (ii) we propose a novel self-supervised co-training scheme to improve the popular infoNCE loss, exploiting the complementary information from different views, RGB streams and optical flow, of the same data source by using one view to obtain positive class samples for the other; (iii) we thoroughly evaluate the quality of the learnt representation on two different downstream tasks: action recognition and video retrieval. In both cases, the proposed approach demonstrates state-of-the-art or comparable performance with other self-supervised approaches, whilst being significantly more efficient to train, i.e. requiring far less training data to achieve similar performance.

[Gradient Estimation with Stochastic Softmax Tricks](#)

- Max Paulus, Dami Choi, Daniel Tarlow, Andreas Krause, Chris J. Maddison
- abstract@[open-review](#): The Gumbel-Max trick is the basis of many relaxed gradient estimators. These estimators are easy to implement and low variance, but the goal of scaling them comprehensively to large combinatorial distributions is still outstanding. Working within the perturbation model framework, we introduce stochastic softmax tricks, which generalize the Gumbel-Softmax trick to combinatorial spaces. Our framework is a unified perspective on existing relaxed estimators for perturbation models, and it contains many novel relaxations. We design structured relaxations for subset selection, spanning trees, arborescences, and others. When compared to less structured baselines, we find that stochastic softmax tricks can be used to train latent variable models that perform better and discover more latent structure.

[Meta-Learning Requires Meta-Augmentation](#)

- Janarthanan Rajendran, Alexander Irpan, Eric Jang
- abstract@[open-review](#): Meta-learning algorithms aim to learn two components: a model that predicts targets for a task, and a base learner that updates that model when given examples from a new task. This additional level of learning can be powerful, but it also creates another potential source of overfitting, since we can now overfit in either the model or the base learner. We describe both of these forms of meta-learning overfitting, and demonstrate that they appear experimentally in common meta-learning benchmarks. We introduce an information-theoretic framework of meta-augmentation, whereby adding randomness discourages the base learner and model from learning trivial solutions that do not generalize to new tasks. We demonstrate that meta-augmentation produces large complementary benefits to recently proposed meta-regularization techniques.

[SLIP: Learning to predict in unknown dynamical systems with long-term memory](#)

- Paria Rashidinejad, Jiantao Jiao, Stuart Russell
- abstract@[open-review](#): We present an efficient and practical (polynomial time) algorithm for online prediction in unknown and partially observed linear dynamical systems (LDS) under stochastic noise. When the system parameters are known, the optimal linear predictor is the Kalman filter. However, in unknown systems, the performance of existing predictive models is poor in important classes of LDS that are only marginally stable and exhibit long-term forecast memory. We tackle this problem by bounding the generalized Kolmogorov width of the Kalman filter coefficient set. This motivates the design of an algorithm, which we call spectral LDS improper predictor (SLIP), based on conducting a tight convex relaxation of the Kalman predictive model via spectral methods. We provide a finite-sample analysis, showing that our algorithm competes with the Kalman filter in hindsight with only logarithmic regret. Our regret analysis relies on Mendelsonâ€™s small-ball method, providing sharp error bounds without concentration, boundedness, or exponential forgetting assumptions. Empirical evaluations demonstrate that SLIP outperforms state-of-the-art methods in LDS prediction. Our theoretical and experimental results shed light on the conditions required for efficient probably approximately correct (PAC) learning of the Kalman filter from partially observed data.

[Improving GAN Training with Probability Ratio Clipping and Sample Reweighting](#)

- Yue Wu, Pan Zhou, Andrew G. Wilson, Eric Xing, Zhiteng Hu
- abstract@[open-review](#): Despite success on a wide range of problems related to vision, generative adversarial networks (GANs) often suffer from inferior performance due to unstable training, especially for text generation. To solve this issue, we propose a new variational GAN training framework which enjoys superior training stability. Our approach is inspired by a connection of GANs and reinforcement learning under a variational perspective. The connection leads to (1) probability ratio clipping that regularizes generator training to prevent excessively large updates, and (2) a sample re-weighting mechanism that improves discriminator training by downplaying bad-quality fake samples. Moreover, our variational GAN framework can provably overcome the training issue in many GANs that an optimal discriminator cannot provide any informative gradient to training generator. By plugging the training approach in diverse state-of-the-art GAN architectures, we obtain significantly improved performance over a range of tasks, including text generation, text style transfer, and image generation.

[Bayesian Bits: Unifying Quantization and Pruning](#)

- Mart van Baalen, Christos Louizos, Markus Nagel, Rana Ali Amjad, Ying Wang, Tijmen Blankevoort, Max Welling
- abstract@[open-review](#): We introduce Bayesian Bits, a practical method for joint mixed precision quantization and pruning through gradient based optimization. Bayesian Bits employs a novel decomposition of the quantization operation, which sequentially considers doubling the bit width. At each new bit width, the residual error between the full precision value and the previously rounded value is quantized. We then decide whether or not to add this quantized residual error for a higher effective bit width and lower quantization noise. By starting with a power-of-two bit width, this decomposition will always produce hardware-friendly configurations, and through an additional 0-bit option, serves as a unified view of pruning and quantization. Bayesian Bits then introduces learnable stochastic gates, which collectively control the bit width of the given tensor. As a result, we can obtain low bit solutions by performing approximate inference over the gates, with prior distributions that encourage most of them to be switched off. We experimentally validate our proposed method on several benchmark datasets and show that we can learn pruned, mixed precision networks that provide a better trade-off between accuracy and efficiency than their static bit width equivalents.

[On Testing of Samplers](#)

- Kuldeep S Meel, Yash Pralhad Pote, Sourav Chakraborty
- abstract@[open-review](#): Given a set of items F and a weight function $W: F \rightarrow (0,1)$, the problem of sampling seeks to sample an item proportional to its weight. Sampling is a fundamental problem in machine learning. The daunting computational complexity of sampling with formal guarantees leads designers to propose heuristics-based techniques for which no rigorous theoretical analysis exists to quantify the quality of the generated distributions. This poses a challenge in designing a testing methodology to test whether a sampler under test generates samples according to a given distribution. Only recently, Chakraborty and Meel (2019) designed the first scalable verifier, called Barbarik, for samplers in the special case when the weight function W is constant, that is, when the sampler is supposed to sample uniformly from F . The techniques in Barbarik, however, fail to handle general weight functions. The primary contribution of this paper is an affirmative answer to the above challenge: motivated by Barbarik, but using different techniques and analysis, we design Barbarik2, an algorithm to test whether the distribution generated by a sampler is epsilon-close or eta-far from any target distribution. In contrast to black-box sampling techniques that require a number of samples proportional to $|F|$, Barbarik2 requires only $\tilde{O}(\text{Tilt}(W, F)^2/\eta^2 - 6^*\epsilon^3)$ samples, where the Tilt is the maximum ratio of weights of two points in F . Barbarik2 can handle any arbitrary weight function. We present a prototype implementation of Barbarik2 and use it to test three state-of-the-art samplers.

[Gaussian Process Bandit Optimization of the Thermodynamic Variational Objective](#)

- Vu Nguyen, Vaden Masrani, Rob Brekelmans, Michael Osborne, Frank Wood
- abstract@[open-review](#): Achieving the full promise of the Thermodynamic Variational Objective (TVO), a recently proposed variational inference objective that lower-bounds the log evidence via one-dimensional Riemann integration, requires choosing a ``schedule'' of sorted discretization points. This paper introduces a bespoke Gaussian process bandit optimization method for automatically choosing these points. Our approach not only automates their one-time selection, but also dynamically adapts their positions over the course of optimization, leading to improved model learning and inference. We provide theoretical guarantees that our bandit optimization converges to the regret-minimizing choice of integration points. Empirical validation of our algorithm is provided in terms of improved learning and inference in Variational Autoencoders and sigmoid belief networks.

[MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers](#)

- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, Ming Zhou
- abstract@[open-review](#): Pre-trained language models (e.g., BERT (Devlin et al., 2018) and its variants) have achieved remarkable success in varieties of NLP tasks. However, these models usually consist of hundreds of millions of parameters which brings challenges for fine-tuning and online serving in real-life applications due to latency and capacity constraints. In this work, we present a simple and effective approach to compress large Transformer (Vaswani et al., 2017) based pre-trained models, termed as deep self-attention distillation. The small model (student) is trained by deeply mimicking the self-attention module, which plays a vital role in Transformer networks, of the large model (teacher). Specifically, we propose distilling the self-attention module of the last Transformer layer of the teacher, which is effective and flexible for the student. Furthermore, we introduce the scaled dot-product between values in the self-attention module as the new deep self-attention knowledge, in addition to the attention distributions (i.e., the scaled dot-product of queries and keys) that have been used in existing works. Moreover, we show that introducing a teacher assistant (Mirzadeh et al., 2019) also helps the distillation of large pre-trained Transformer models. Experimental results demonstrate that our monolingual model outperforms state-of-the-art baselines in different parameter size of student models. In particular, it retains more than 99% accuracy on SQuAD 2.0 and several GLUE benchmark tasks using 50% of the Transformer parameters and computations of the teacher model. We also obtain competitive results in applying deep self-attention distillation to multilingual pre-trained models.

[Optimal Epoch Stochastic Gradient Descent Ascent Methods for Min-Max Optimization](#)

- Yan Yan, Yi Xu, Qihang Lin, Wei Liu, Tianbao Yang
- abstract@[open-review](#): Epoch gradient descent method (a.k.a. Epoch-GD) proposed by (Hazan and Kale, 2011) was deemed a breakthrough for stochastic strongly convex minimization, which achieves the optimal convergence rate of $O(1/T)$ with T iterative updates for the objective gap. However, its extension to solving stochastic min-max problems with strong convexity and strong concavity still remains open, and it is still unclear whether a fast rate of $O(1/T)$ for the duality gap is achievable for stochastic min-max optimization under strong convexity and strong concavity. Although some recent studies have proposed stochastic algorithms with fast convergence rates for min-max problems, they require additional assumptions about the problem, e.g., smoothness, bi-linear structure, etc. In this paper, we bridge this gap by providing a sharp analysis of epoch-wise stochastic gradient descent ascent method (referred to as Epoch-GDA) for solving strongly convex strongly concave (SCSC) min-max problems, without imposing any additional assumption about smoothness or the function's structure. To the best of our knowledge, our result is the first one that shows Epoch-GDA can achieve the optimal rate of $O(1/T)$ for the duality gap of general SCSC min-max problems. We emphasize that such generalization of Epoch-GD for strongly convex minimization problems to Epoch-GDA for SCSC min-max problems is non-trivial and requires novel technical analysis. Moreover, we notice that the key lemma can also be used for proving the convergence of Epoch-GDA for weakly-convex strongly-concave min-max problems, leading to a nearly optimal complexity without resorting to smoothness or other structural conditions.

[Woodbury Transformations for Deep Generative Flows](#)

- You Lu, Bert Huang
- abstract@[open-review](#): Normalizing flows are deep generative models that allow efficient likelihood calculation and sampling. The core requirement for this advantage is that they are constructed using functions that can be efficiently inverted and for which the determinant of the function's Jacobian can be efficiently computed. Researchers have introduced various such flow operations, but few of these allow rich interactions among variables without incurring significant computational costs. In this paper, we introduce Woodbury transformations, which achieve efficient invertibility via the Woodbury matrix identity and efficient determinant calculation via Sylvester's determinant identity. In contrast with other operations used in state-of-the-art normalizing flows, Woodbury transformations enable (1) high-dimensional interactions, (2) efficient sampling, and (3) efficient likelihood evaluation. Other similar operations, such as 1x1 convolutions, emerging convolutions, or periodic convolutions allow at most two of these three advantages. In our experiments on multiple image datasets, we find that Woodbury transformations allow learning of higher-likelihood models than other flow architectures while still enjoying their efficiency advantages.

[Graph Contrastive Learning with Augmentations](#)

- Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, Yang Shen
- abstract@[open-review](#): Generalizable, transferable, and robust representation learning on graph-structured data remains a challenge for current graph neural networks (GNNs). Unlike what has been developed for convolutional neural networks (CNNs) for image data, self-supervised learning and pre-training are less explored for GNNs. In this paper, we propose a graph contrastive learning (GraphCL) framework for learning unsupervised representations of graph data. We first design four types of graph augmentations to incorporate various priors. We then systematically study the impact of various combinations of graph augmentations on multiple datasets, in four different settings: semi-supervised, unsupervised, and transfer learning as well as adversarial attacks. The results show that, even without tuning augmentation extents nor using sophisticated GNN architectures, our GraphCL framework can produce graph representations of similar or better generalizability, transferability, and robustness compared to state-of-the-art methods. We also investigate the impact of parameterized graph augmentation extents and patterns, and observe further performance gains in preliminary experiments. Our codes are available at <https://github.com/Shen-Lab/GraphCL>.

[Gradient Surgery for Multi-Task Learning](#)

- Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, Chelsea Finn
- abstract@[open-review](#): While deep learning and deep reinforcement learning (RL) systems have demonstrated impressive results in domains such as image classification, game playing, and robotic control, data efficiency remains a major challenge. Multi-task learning has emerged as a promising approach for sharing structure across multiple tasks to enable more efficient learning. However, the multi-task setting presents a number of optimization challenges, making it difficult to realize large efficiency gains compared to learning tasks independently. The reasons why multi-task learning is so challenging compared to single-task learning are not fully understood. In this work, we identify a set of three conditions of the multi-task optimization landscape that cause detrimental gradient interference, and develop a simple yet general approach for avoiding such interference between task gradients. We propose a form of gradient surgery that projects a task's gradient onto the normal plane of the gradient of any other task that has a gradient. On a series of challenging multi-task supervised and multi-task RL problems, this approach leads to substantial gains in efficiency and performance. Further, it is model-agnostic and can be combined with previously-proposed multi-task architectures for enhanced performance.

[Bayesian Probabilistic Numerical Integration with Tree-Based Models](#)

- Harrison Zhu, Xing Liu, Ruya Kang, Zhichao Shen, Seth Flaxman, Francois-Xavier Briol
- abstract@[open-review](#): Bayesian quadrature (BQ) is a method for solving numerical integration problems in a Bayesian manner, which allows users to quantify their uncertainty about the solution. The standard approach to BQ is based on a Gaussian process (GP) approximation of the integrand. As a result, BQ is inherently limited to cases where GP approximations can be done in an efficient manner, thus often prohibiting very high-dimensional or non-smooth target functions. This paper proposes to tackle this issue with a new Bayesian numerical integration algorithm based on Bayesian Additive Regression Trees (BART) priors, which we call BART-Int. BART priors are easy to tune and well-suited for discontinuous functions. We demonstrate that they also lend themselves naturally to a sequential design setting and that explicit convergence rates can be obtained in a variety of settings. The advantages and disadvantages of this new methodology are highlighted on a set of benchmark tests including the Genz functions, on a rare-event simulation problem and on a Bayesian survey design problem.

[Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the Neural Tangent Kernel](#)

- Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M. Roy, Surya Ganguli
- abstract@[open-review](#): In suitably initialized wide networks, small learning rates transform deep neural networks (DNNs) into neural tangent kernel (NTK) machines, whose training dynamics is well-approximated by a linear weight expansion of the network at initialization. Standard training, however, diverges from its linearization in ways that are poorly understood. We study the relationship between the training dynamics of nonlinear deep networks, the geometry of the loss landscape, and the time evolution of a data-dependent NTK. We do so through a large-scale phenomenological analysis of training, synthesizing diverse measures characterizing loss landscape geometry and NTK dynamics. In multiple neural architectures and datasets, we find these diverse measures evolve in a highly correlated manner, revealing a universal picture of the deep learning process. In this picture, deep network training exhibits a highly chaotic rapid initial transient that within 2 to 3 epochs determines the final linearly connected basin of low loss containing the end point of training. During this chaotic transient, the NTK changes rapidly, learning useful features from the training data that enables it to outperform the standard initial NTK by a factor of 3 in less than 3 to 4 epochs. After this rapid chaotic transient, the NTK changes at constant velocity, and its performance matches that of full network training in 15% to 45% of training time. Overall, our analysis reveals a striking correlation between a diverse set of metrics over training time, governed by a rapid chaotic to stable transition in the first few epochs, that together poses challenges and opportunities for the development of more accurate theories of deep learning.

[Graph Meta Learning via Local Subgraphs](#)

- Kexin Huang, Marinka Zitnik
- abstract@[open-review](#): Prevailing methods for graphs require abundant label and edge information for learning. When data for a new task are scarce, meta-learning can learn from prior experiences and form much-needed inductive biases for fast adaption to new tasks. Here, we introduce G-Meta, a novel meta-learning algorithm for graphs. G-Meta uses local subgraphs to transfer subgraph-specific information and learn transferable knowledge faster via meta gradients. G-Meta learns how to quickly adapt to a new task using only a handful of nodes or edges in the new task and does so by learning from data points in other graphs or related, albeit disjoint label sets. G-Meta is theoretically justified as we show that the evidence for a prediction can be found in the local subgraph surrounding the target node or edge. Experiments on seven datasets and nine baseline methods show that G-Meta outperforms existing methods by up to 16.3%. Unlike previous methods, G-Meta successfully learns in challenging, few-shot learning settings that require generalization to completely new graphs and never-before-seen labels. Finally, G-Meta scales to large graphs, which we demonstrate on a new Tree-of-Life dataset comprising of 1,840 graphs, a two-orders of magnitude increase in the number of graphs used in prior work.

[Stochastic Deep Gaussian Processes over Graphs](#)

- Naiqi Li, Wenjie Li, Jifeng Sun, Yinghua Gao, Yong Jiang, Shu-Tao Xia
- abstract@[open-review](#): In this paper we propose Stochastic Deep Gaussian Processes over Graphs (DGP), which are deep structure models that learn the mappings between input and output signals in graph domains. The approximate posterior distributions of the latent variables are derived with variational inference, and the evidence lower bound is evaluated and optimized by the proposed recursive sampling scheme. The Bayesian non-parametric nature of our model allows it to resist overfitting, while the expressive deep structure grants it the potential to learn complex relations. Extensive experiments demonstrate that our method achieves superior performances in both small size (< 50) and large size (> 35,000) datasets. We show that DGP outperforms another Gaussian-based approach, and is competitive to a state-of-the-art method in the challenging task of traffic flow prediction. Our model is also capable of capturing uncertainties in a mathematical principled way and automatically discovering which vertices and features are relevant to the prediction.

[Bayesian Causal Structural Learning with Zero-Inflated Poisson Bayesian Networks](#)

- Junsouk Choi, Robert Chapkin, Yang Ni
- abstract@[open-review](#): Multivariate zero-inflated count data arise in a wide range of areas such as economics, social sciences, and biology. To infer causal relationships in zero-inflated count data, we propose a new zero-inflated Poisson Bayesian network (ZIPBN) model. We show that the proposed ZIPBN is identifiable with cross-sectional data. The proof is based on the well-known characterization of Markov equivalence class which is applicable to other distribution families. For causal structural learning, we introduce a fully Bayesian inference approach which exploits the parallel tempering Markov chain Monte Carlo algorithm to efficiently explore the multi-modal network space. We demonstrate the utility of the proposed ZIPBN in causal discoveries for zero-inflated count data by simulation studies with comparison to alternative Bayesian network methods. Additionally, real single-cell RNA-sequencing data with known causal relationships will be used to assess the capability of ZIPBN for discovering causal relationships in real-world problems.

[Evaluating Attribution for Graph Neural Networks](#)

- Benjamin Sanchez-Lengeling, Jennifer Wei, Brian Lee, Emily Reif, Peter Wang, Wesley Qian, Kevin McCloskey, Lucy Colwell, Alexander Wiltschko
- abstract@[open-review](#): Interpretability of machine learning models is critical to scientific understanding, AI safety, as well as debugging. Attribution is one approach to interpretability, which highlights input dimensions that are influential to a neural network's prediction. Evaluation of these methods is largely qualitative for image and text models, because acquiring ground truth attributions requires expensive and unreliable human judgment. Attribution has been little studied for graph neural networks (GNNs), a model class of growing importance that makes predictions on arbitrarily-sized graphs. In this work we adapt commonly-used attribution methods for GNNs and quantitatively evaluate them using computable ground-truths that are objective and challenging to learn. We make concrete recommendations for which attribution methods to use, and provide the data and code for our benchmarking suite. Rigorous and open source benchmarking of attribution methods in graphs could enable new methods development and broader use of attribution in real-world ML tasks.

[On Second Order Behaviour in Augmented Neural ODEs](#)

- Alexander Norcliffe, Cristian Bodnar, Ben Day, Nikola Simidjievski, Pietro Li³
- abstract@[open-review](#): Neural Ordinary Differential Equations (NODEs) are a new class of models that transform data continuously through infinite-depth architectures. The continuous nature of NODEs has made them particularly suitable for learning the dynamics of complex physical systems. While previous work has mostly been focused on first order ODEs, the dynamics of many systems, especially in classical physics, are governed by second order laws. In this work, we consider Second Order Neural ODEs (SONODEs). We show how the adjoint sensitivity method can be extended to SONODEs and

prove that the optimisation of a first order coupled ODE is equivalent and computationally more efficient. Furthermore, we extend the theoretical understanding of the broader class of Augmented NODEs (ANODEs) by showing they can also learn higher order dynamics with a minimal number of augmented dimensions, but at the cost of interpretability. This indicates that the advantages of ANODEs go beyond the extra space offered by the augmented dimensions, as originally thought. Finally, we compare SONODEs and ANODEs on synthetic and real dynamical systems and demonstrate that the inductive biases of the former generally result in faster training and better performance.

[Neuron Shapley: Discovering the Responsible Neurons](#)

- Amirata Ghorbani, James Y. Zou
- abstract@[open-review](#): We develop Neuron Shapley as a new framework to quantify the contribution of individual neurons to the prediction and performance of a deep network. By accounting for interactions across neurons, Neuron Shapley is more effective in identifying important filters compared to common approaches based on activation patterns. Interestingly, removing just 30 filters with the highest Shapley scores effectively destroys the prediction accuracy of Inception-v3 on ImageNet. Visualization of these few critical filters provides insights into how the network functions. Neuron Shapley is a flexible framework and can be applied to identify responsible neurons in many tasks. We illustrate additional applications of identifying filters that are responsible for biased prediction in facial recognition and filters that are vulnerable to adversarial attacks. Removing these filters is a quick way to repair models. Computing exact Shapley values is computationally infeasible and therefore sampling-based approximations are used in practice. We introduce a new multi-armed bandit algorithm that is able to efficiently detect neurons with the largest Shapley value orders of magnitude faster than existing Shapley value approximation methods.

[Stochastic Normalizing Flows](#)

- Hao Wu, Jonas Kähler, Frank Noe
- abstract@[open-review](#): The sampling of probability distributions specified up to a normalization constant is an important problem in both machine learning and statistical mechanics. While classical stochastic sampling methods such as Markov Chain Monte Carlo (MCMC) or Langevin Dynamics (LD) can suffer from slow mixing times there is a growing interest in using normalizing flows in order to learn the transformation of a simple prior distribution to the given target distribution. Here we propose a generalized and combined approach to sample target densities: Stochastic Normalizing Flows (SNF) – an arbitrary sequence of deterministic invertible functions and stochastic sampling blocks. We show that stochasticity overcomes expressivity limitations of normalizing flows resulting from the invertibility constraint, whereas trainable transformations between sampling steps improve efficiency of pure MCMC/LD along the flow. By invoking ideas from non-equilibrium statistical mechanics we derive an efficient training procedure by which both the sampler's and the flow's parameters can be optimized end-to-end, and by which we can compute exact importance weights without having to marginalize out the randomness of the stochastic blocks. We illustrate the representational power, sampling efficiency and asymptotic correctness of SNFs on several benchmarks including applications to sampling molecular systems in equilibrium.

[GPU-Accelerated Primal Learning for Extremely Fast Large-Scale Classification](#)

- John T. Halloran, David M. Rocke
- abstract@[open-review](#): One of the most efficient methods to solve L2 -regularized primal problems, such as logistic regression and linear support vector machine (SVM) classification, is the widely used trust region Newton algorithm, TRON. While TRON has recently been shown to enjoy substantial speedups on shared-memory multi-core systems, exploiting graphical processing units (GPUs) to speed up the method is significantly more difficult, owing to the highly complex and heavily sequential nature of the algorithm. In this work, we show that using judicious GPU-optimization principles, TRON training time for different losses and feature representations may be drastically reduced. For sparse feature sets, we show that using GPUs to train logistic regression classifiers in LIBLINEAR is up to an order-of-magnitude faster than solely using multithreading. For dense feature sets – which impose far more stringent memory constraints – we show that GPUs substantially reduce the lengthy SVM learning times required for state-of-the-art proteomics analysis, leading to dramatic improvements over recently proposed speedups. Furthermore, we show how GPU speedups may be mixed with multithreading to enable such speedups when the dataset is too large for GPU memory requirements; on a massive dense proteomics dataset of nearly a quarter-billion data instances, these mixed-architecture speedups reduce SVM analysis time from over half a week to less than a single day while using limited GPU memory.

[Random Reshuffling is Not Always Better](#)

- Christopher M. De Sa
- abstract@[open-review](#): Many learning algorithms, such as stochastic gradient descent, are affected by the order in which training examples are used. It is often observed that sampling the training examples without-replacement, also known as random reshuffling, causes learning algorithms to converge faster. We give a counterexample to the Operator Inequality of Noncommutative Arithmetic and Geometric Means, a longstanding conjecture that relates to the performance of random reshuffling in learning algorithms (Recht and Röder, "Toward a noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences," COLT 2012). We use this to give an example of a learning task and algorithm for which with-replacement random sampling actually outperforms random reshuffling.

[Model Agnostic Multilevel Explanations](#)

- Karthikeyan Natesan Ramamurthy, Bhanukiran Vinzamuri, Yunfeng Zhang, Amit Dhurandhar
- abstract@[open-review](#): In recent years, post-hoc local instance-level and global dataset-level explainability of black-box models has received a lot of attention. Lesser attention has been given to obtaining insights at intermediate or group levels, which is a need outlined in recent works that study the challenges in realizing the guidelines in the General Data Protection Regulation (GDPR). In this paper, we propose a meta-method that, given a typical local explainability method, can build a multilevel explanation tree. The leaves of this tree correspond to local explanations, the root corresponds to global explanation, and intermediate levels correspond to explanations for groups of data points that it automatically clusters. The method can also leverage side information, where users can specify points for which they may want the explanations to be similar. We argue that such a multilevel structure can also be an effective form of communication, where one could obtain few explanations that characterize the entire dataset by considering an appropriate level in our explanation tree. Explanations for novel test points can be cost-efficiently obtained by associating them with the closest training points. When the local explainability technique is generalized additive (viz. LIME, GAMs), we develop fast approximate algorithm for building the multilevel tree and study its convergence behavior. We show that we produce high fidelity sparse explanations on several public datasets and also validate the effectiveness of the proposed technique based on two human studies -- one with experts and the other with non-expert users -- on real world datasets.

[NeuMiss networks: differentiable programming for supervised learning with missing values.](#)

- Marine Le Morvan, Julie Josse, Thomas Moreau, Erwan Scornet, Gael Varoquaux
- abstract@[open-review](#): The presence of missing values makes supervised learning much more challenging. Indeed, previous work has shown that even when the response is a linear function of the complete data, the optimal predictor is a complex function of the observed entries and the missingness indicator. As a result, the computational or sample complexities of consistent approaches depend on the number of missing patterns, which can be exponential in the number of dimensions. In this work, we derive the analytical form of the optimal predictor under a linearity assumption and various missing data mechanisms including Missing at Random (MAR) and self-masking (Missing Not At Random). Based on a Neumann-series approximation of the optimal predictor, we propose a new principled architecture, named NeuMiss networks. Their originality and strength come from the use of a new

type of non-linearity: the multiplication by the missingness indicator. We provide an upper bound on the Bayes risk of NeuMiss networks, and show that they have good predictive accuracy with both a number of parameters and a computational complexity independent of the number of missing data patterns. As a result they scale well to problems with many features, and remain statistically efficient for medium-sized samples. Moreover, we show that, contrary to procedures using EM or imputation, they are robust to the missing data mechanism, including difficult MNAR settings such as self-masking.

[Revisiting Parameter Sharing for Automatic Neural Channel Number Search](#)

- Jiaxing Wang, Haoli Bai, Jiaxiang Wu, Xupeng Shi, Junzhou Huang, Irwin King, Michael Lyu, Jian Cheng
- abstract@[open-review](#): Recent advances in neural architecture search inspire many channel number search algorithms~(CNS) for convolutional neural networks. To improve searching efficiency, parameter sharing is widely applied, which reuses parameters among different channel configurations. Nevertheless, it is unclear how parameter sharing affects the searching process. In this paper, we aim at providing a better understanding and exploitation of parameter sharing for CNS. Specifically, we propose affine parameter sharing~(APS) as a general formulation to unify and quantitatively analyze existing channel search algorithms. It is found that with parameter sharing, weight updates of one architecture can simultaneously benefit other candidates. However, it also results in less confidence in choosing good architectures. We thus propose a new strategy of parameter sharing towards a better balance between training efficiency and architecture discrimination. Extensive analysis and experiments demonstrate the superiority of the proposed strategy in channel configuration against many state-of-the-art counterparts on benchmark datasets.

[Differentially-Private Federated Linear Bandits](#)

- Abhimanyu Dubey, Alex `Sandy' Pentland
- abstract@[open-review](#): The rapid proliferation of decentralized learning systems mandates the need for differentially-private cooperative learning. In this paper, we study this in context of the contextual linear bandit: we consider a collection of agents cooperating to solve a common contextual bandit, while ensuring that their communication remains private. For this problem, we devise FedUCB, a multiagent private algorithm for both centralized and decentralized (peer-to-peer) federated learning. We provide a rigorous technical analysis of its utility in terms of regret, improving several results in cooperative bandit learning, and provide rigorous privacy guarantees as well. Our algorithms provide competitive performance both in terms of pseudoregret bounds and empirical benchmark performance in various multi-agent settings.

[Is Plug-in Solver Sample-Efficient for Feature-based Reinforcement Learning?](#)

- Qiwen Cui, Lin Yang
- abstract@[open-review](#): It is believed that a model-based approach for reinforcement learning (RL) is the key to reduce sample complexity. However, the understanding of the sample optimality of model-based RL is still largely missing, even for the linear case. This work considers sample complexity of finding an $\$\\epsilon$ -optimal policy in a Markov decision process (MDP) that admits a linear additive feature representation, given only access to a generative model. We solve this problem via a plug-in solver approach, which builds an empirical model and plans in this empirical model via an arbitrary plug-in solver. We prove that under the anchor-state assumption, which implies implicit non-negativity in the feature space, the minimax sample complexity of finding an $\$\\epsilon$ -optimal policy in a $\$\\gamma$ -discounted MDP is $\$O(K/(1-\\gamma)^3\\epsilon^2)$, which only depends on the dimensionality $\$K\$$ of the feature space and has no dependence on the state or action space. We further extend our results to a relaxed setting where anchor-states may not exist and show that a plug-in approach can be sample efficient as well, providing a flexible approach to design model-based algorithms for RL.

[Learning Physical Graph Representations from Visual Scenes](#)

- Daniel Bear, Chaofei Fan, Damian Mrowca, Yunzhu Li, Seth Alter, Aran Nayebi, Jeremy Schwartz, Li F. Fei-Fei, Jiajun Wu, Josh Tenenbaum, Daniel L. Yamins
- abstract@[open-review](#): Convolutional Neural Networks (CNNs) have proved exceptional at learning representations for visual object categorization. However, CNNs do not explicitly encode objects, parts, and their physical properties, which has limited CNNs' success on tasks that require structured understanding of visual scenes. To overcome these limitations, we introduce the idea of ``Physical Scene Graphs'' (PSGs), which represent scenes as hierarchical graphs, with nodes in the hierarchy corresponding intuitively to object parts at different scales, and edges to physical connections between parts. Bound to each node is a vector of latent attributes that intuitively represent object properties such as surface shape and texture. We also describe PSGNet, a network architecture that learns to extract PSGs by reconstructing scenes through a PSG-structured bottleneck. PSGNet augments standard CNNs by including: recurrent feedback connections to combine low and high-level image information; graph pooling and vectorization operations that convert spatially-uniform feature maps into object-centric graph structures; and perceptual grouping principles to encourage the identification of meaningful scene elements. We show that PSGNet outperforms alternative self-supervised scene representation algorithms at scene segmentation tasks, especially on complex real-world images, and generalizes well to unseen object types and scene arrangements. PSGNet is also able learn from physical motion, enhancing scene estimates even for static images. We present a series of ablation studies illustrating the importance of each component of the PSGNet architecture, analyses showing that learned latent attributes capture intuitive scene properties, and illustrate the use of PSGs for compositional scene inference.

[Deep Graph Pose: a semi-supervised deep graphical model for improved animal pose tracking](#)

- Anqi Wu, Estefany Kelly Buchanan, Matthew Whiteway, Michael Schartner, Guido Meijer, Jean-Paul Noel, Erica Rodriguez, Claire Everett, Amy Norovich, Evan Schaffer, Neeli Mishra, C. Daniel Salzman, Dora Angelaki, AndrÃ©s Bendesky, The International Brain Laboratory The International Brain Laboratory, John P. Cunningham, Liam Paninski
- abstract@[open-review](#): Noninvasive behavioral tracking of animals is crucial for many scientific investigations. Recent transfer learning approaches for behavioral tracking have considerably advanced the state of the art. Typically these methods treat each video frame and each object to be tracked independently. In this work, we improve on these methods (particularly in the regime of few training labels) by leveraging the rich spatiotemporal structures pervasive in behavioral video --- specifically, the spatial statistics imposed by physical constraints (e.g., paw to elbow distance), and the temporal statistics imposed by smoothness from frame to frame. We propose a probabilistic graphical model built on top of deep neural networks, Deep Graph Pose (DGP), to leverage these useful spatial and temporal constraints, and develop an efficient structured variational approach to perform inference in this model. The resulting semi-supervised model exploits both labeled and unlabeled frames to achieve significantly more accurate and robust tracking while requiring users to label fewer training frames. In turn, these tracking improvements enhance performance on downstream applications, including robust unsupervised segmentation of behavioral syllables," and estimation of interpretable disentangled" low-dimensional representations of the full behavioral video. Open source code is available at \href{\CodeLink}{<https://github.com/paninski-lab/deepgraphpose>}.

[Meta-learning from Tasks with Heterogeneous Attribute Spaces](#)

- Tomoharu Iwata, Atsutoshi Kumagai
- abstract@[open-review](#): We propose a heterogeneous meta-learning method that trains a model on tasks with various attribute spaces, such that it can solve unseen tasks whose attribute spaces are different from the training tasks given a few labeled instances. Although many meta-learning methods have been proposed, they assume that all training and target tasks share the same attribute space, and they are inapplicable when attribute sizes are different across tasks. Our model infers latent representations of each attribute and each response from a few labeled instances using an inference network. Then, responses of unlabeled instances are predicted with the inferred representations using a prediction network. The attribute and response representations

enable us to make predictions based on the task-specific properties of attributes and responses even when attribute and response sizes are different across tasks. In our experiments with synthetic datasets and 59 datasets in OpenML, we demonstrate that our proposed method can predict the responses given a few labeled instances in new tasks after being trained with tasks with heterogeneous attribute spaces.

Estimating decision tree learnability with polylogarithmic sample complexity

- Guy Blanc, Neha Gupta, Jane Lange, Li-Yang Tan
- abstract@[open-review](#): We show that top-down decision tree learning heuristics (such as ID3, C4.5, and CART) are amenable to highly efficient \{\\sl learnability estimation\}: for monotone target functions, the error of the decision tree hypothesis constructed by these heuristics can be estimated with \{\\sl polylogarithmically\} many labeled examples, exponentially smaller than the number necessary to run these heuristics, and indeed, exponentially smaller than information-theoretic minimum required to learn a good decision tree. This adds to a small but growing list of fundamental learning algorithms that have been shown to be amenable to learnability estimation.

En route to this result, we design and analyze sample-efficient \{\\sl minibatch\} versions of top-down decision tree learning heuristics and show that they achieve the same provable guarantees as the full-batch versions. We further give ``active local'' versions of these heuristics: given a test point x^* , we show how the label $T(x^*)$ of the decision tree hypothesis T can be computed with polylogarithmically many labeled examples, exponentially smaller than the number necessary to learn T .

Sparse Symplectically Integrated Neural Networks

- Daniel DiPietro, Shiying Xiong, Bo Zhu
- abstract@[open-review](#): We introduce Sparse Symplectically Integrated Neural Networks (SSINNs), a novel model for learning Hamiltonian dynamical systems from data. SSINNs combine fourth-order symplectic integration with a learned parameterization of the Hamiltonian obtained using sparse regression through a mathematically elegant function space. This allows for interpretable models that incorporate symplectic inductive biases and have low memory requirements. We evaluate SSINNs on four classical Hamiltonian dynamical problems: the H̄non-Heiles system, nonlinearly coupled oscillators, a multi-particle mass-spring system, and a pendulum system. Our results demonstrate promise in both system prediction and conservation of energy, often outperforming the current state-of-the-art black-box prediction techniques by an order of magnitude. Further, SSINNs successfully converge to true governing equations from highly limited and noisy data, demonstrating potential applicability in the discovery of new physical governing equations.

Continuous Object Representation Networks: Novel View Synthesis without Target View Supervision

- Nicolai Hani, Selim Engin, Jun-Jee Chao, Volkan Isler
- abstract@[open-review](#): Novel View Synthesis (NVS) is concerned with synthesizing views under camera viewpoint transformations from one or multiple input images. NVS requires explicit reasoning about 3D object structure and unseen parts of the scene to synthesize convincing results. As a result, current approaches typically rely on supervised training with either ground truth 3D models or multiple target images. We propose Continuous Object Representation Networks (CORN), a conditional architecture that encodes an input image's geometry and appearance that map to a 3D consistent scene representation. We can train CORN with only two source images per object by combining our model with a neural renderer. A key feature of CORN is that it requires no ground truth 3D models or target view supervision. Regardless, CORN performs well on challenging tasks such as novel view synthesis and single-view 3D reconstruction and achieves performance comparable to state-of-the-art approaches that use direct supervision. For up-to-date information, data, and code, please see our project page: <https://nicolaihaeni.github.io/corn/>.

Multimodal Generative Learning Utilizing Jensen-Shannon-Divergence

- Thomas Sutter, Imant Daunhauer, Julia Vogt
- abstract@[open-review](#): Learning from different data types is a long-standing goal in machine learning research, as multiple information sources co-occur when describing natural phenomena. However, existing generative models that approximate a multimodal ELBO rely on difficult or inefficient training schemes to learn a joint distribution and the dependencies between modalities. In this work, we propose a novel, efficient objective function that utilizes the Jensen-Shannon divergence for multiple distributions. It simultaneously approximates the unimodal and joint multimodal posteriors directly via a dynamic prior. In addition, we theoretically prove that the new multimodal JS-divergence (mmJSD) objective optimizes an ELBO. In extensive experiments, we demonstrate the advantage of the proposed mmJSD model compared to previous work in unsupervised, generative learning tasks.

Solver-in-the-Loop: Learning from Differentiable Physics to Interact with Iterative PDE-Solvers

- Kiwon Um, Robert Brand, Yun (Raymond) Fei, Philipp Holl, Nils Thuerey
- abstract@[open-review](#): Finding accurate solutions to partial differential equations (PDEs) is a crucial task in all scientific and engineering disciplines. It has recently been shown that machine learning methods can improve the solution accuracy by correcting for effects not captured by the discretized PDE. We target the problem of reducing numerical errors of iterative PDE solvers and compare different learning approaches for finding complex correction functions. We find that previously used learning approaches are significantly outperformed by methods that integrate the solver into the training loop and thereby allow the model to interact with the PDE during training. This provides the model with realistic input distributions that take previous corrections into account, yielding improvements in accuracy with stable rollouts of several hundred recurrent evaluation steps and surpassing even tailored supervised variants. We highlight the performance of the differentiable physics networks for a wide variety of PDEs, from non-linear advection-diffusion systems to three-dimensional Navier-Stokes flows.

Reinforcement Learning with General Value Function Approximation: Provably Efficient Approach via Bounded Eluder Dimension

- Ruosong Wang, Russ R. Salakhutdinov, Lin Yang
- abstract@[open-review](#): Value function approximation has demonstrated phenomenal empirical success in reinforcement learning (RL). Nevertheless, despite a handful of recent progress on developing theory for RL with linear function approximation, the understanding of \{\\emph{general}\} function approximation schemes largely remains missing. In this paper, we establish the first provably efficient RL algorithm with general value function approximation. We show that if the value functions admit an approximation with a function class \mathcal{F} , our algorithm achieves a regret bound of $\tilde{O}(\mathrm{poly}(dH)\sqrt{T})$ where d is a complexity measure of \mathcal{F} that depends on the eluder dimension [Russo and Van Roy, 2013] and log-covering numbers, H is the planning horizon, and T is the number interactions with the environment. Our theory generalizes the linear MDP assumption to general function classes. Moreover, our algorithm is model-free and provides a framework to justify the effectiveness of algorithms used in practice.

Predicting Training Time Without Training

- Luca Zancato, Alessandro Achille, Avinash Ravichandran, Rahul Bhotika, Stefano Soatto

- abstract@[open-review](#): We tackle the problem of predicting the number of optimization steps that a pre-trained deep network needs to converge to a given value of the loss function. To do so, we leverage the fact that the training dynamics of a deep network during fine-tuning are well approximated by those of a linearized model. This allows us to approximate the training loss and accuracy at any point during training by solving a low-dimensional Stochastic Differential Equation (SDE) in function space. Using this result, we are able to predict the time it takes for Stochastic Gradient Descent (SGD) to fine-tune a model to a given loss without having to perform any training. In our experiments, we are able to predict training time of a ResNet within a 20% error margin on a variety of datasets and hyper-parameters, at a 30 to 45-fold reduction in cost compared to actual training. We also discuss how to further reduce the computational and memory cost of our method, and in particular we show that by exploiting the spectral properties of the gradients' matrix it is possible to predict training time on a large dataset while processing only a subset of the samples.

[How does This Interaction Affect Me? Interpretable Attribution for Feature Interactions](#)

- Michael Tsang, Sirisha Rambhatla, Yan Liu
- abstract@[open-review](#): Machine learning transparency calls for interpretable explanations of how inputs relate to predictions. Feature attribution is a way to analyze the impact of features on predictions. Feature interactions are the contextual dependence between features that jointly impact predictions. There are a number of methods that extract feature interactions in prediction models; however, the methods that assign attributions to interactions are either uninterpretable, model-specific, or non-axiomatic. We propose an interaction attribution and detection framework called Archipelago which addresses these problems and is also scalable in real-world settings. Our experiments on standard annotation labels indicate our approach provides significantly more interpretable explanations than comparable methods, which is important for analyzing the impact of interactions on predictions. We also provide accompanying visualizations of our approach that give new insights into deep neural networks.

[Optimal Adaptive Electrode Selection to Maximize Simultaneously Recorded Neuron Yield](#)

- John Choi, Krishan Kumar, Mohammad Khazali, Katie Wingel, Mahdi Choudhury, Adam S. Charles, Bijan Pesaran
- abstract@[open-review](#): Neural-Matrix style, high-density electrode arrays for brain-machine interfaces (BMIs) and neuroscientific research require the use of multiplexing: Each recording channel can be routed to one of several electrode sites on the array. This capability allows the user to flexibly distribute recording channels to the locations where the most desirable neural signals can be resolved. For example, in the Neuropixel probe, 960 electrodes can be addressed by 384 recording channels. However, currently no adaptive methods exist to use recorded neural data to optimize/customize the electrode selections per recording context. Here, we present an algorithm called classification-based selection (CBS) that optimizes the joint electrode selections for all recording channels so as to maximize isolation quality of detected neurons. We show, in experiments using Neuropixels in non-human primates, that this algorithm yields a similar number of isolated neurons as would be obtained if all electrodes were recorded simultaneously. Neuron counts were 41-85% improved over previously published electrode selection strategies. The neurons isolated from electrodes selected by CBS were a 73% match, by spike timing, to the complete set of recordable neurons around the probe. The electrodes selected by CBS exhibited higher average per-recording-channel signal-to-noise ratio. CBS, and selection optimization in general, could play an important role in development of neurotechnologies for BMI, as signal bandwidth becomes an increasingly limiting factor. Code and experimental data have been made available.

[Neurosymbolic Reinforcement Learning with Formally Verified Exploration](#)

- Greg Anderson, Abhinav Verma, Isil Dillig, Swarat Chaudhuri
- abstract@[open-review](#): We present REVEL, a partially neural reinforcement learning (RL) framework for provably safe exploration in continuous state and action spaces. A key challenge for provably safe deep RL is that repeatedly verifying neural networks within a learning loop is computationally infeasible. We address this challenge using two policy classes: a general, neurosymbolic class with approximate gradients and a more restricted class of symbolic policies that allows efficient verification. Our learning algorithm is a mirror descent over policies: in each iteration, it safely lifts a symbolic policy into the neurosymbolic space, performs safe gradient updates to the resulting policy, and projects the updated policy into the safe symbolic subset, all without requiring explicit verification of neural networks. Our empirical results show that REVEL enforces safe exploration in many scenarios in which Constrained Policy Optimization does not, and that it can discover policies that outperform those learned through prior approaches to verified exploration.

[Wavelet Flow: Fast Training of High Resolution Normalizing Flows](#)

- Jason J. Yu, Konstantinos G. Derpanis, Marcus A. Brubaker
- abstract@[open-review](#): Normalizing flows are a class of probabilistic generative models which allow for both fast density computation and efficient sampling and are effective at modelling complex distributions like images. A drawback among current methods is their significant training cost, sometimes requiring months of GPU training time to achieve state-of-the-art results. This paper introduces Wavelet Flow, a multi-scale, normalizing flow architecture based on wavelets. A Wavelet Flow has an explicit representation of signal scale that inherently includes models of lower resolution signals and conditional generation of higher resolution signals, i.e., super resolution. A major advantage of Wavelet Flow is the ability to construct generative models for high resolution data (e.g., 1024 Å – 1024 images) that are impractical with previous models. Furthermore, Wavelet Flow is competitive with previous normalizing flows in terms of bits per dimension on standard (low resolution) benchmarks while being up to 15Å – faster to train.

[Multi-task Batch Reinforcement Learning with Metric Learning](#)

- Jiachen Li, Quan Vuong, Shuang Liu, Minghua Liu, Kamil Ciosek, Henrik Christensen, Hao Su
- abstract@[open-review](#): We tackle the Multi-task Batch Reinforcement Learning problem. Given multiple datasets collected from different tasks, we train a multi-task policy to perform well in unseen tasks sampled from the same distribution. The task identities of the unseen tasks are not provided. To perform well, the policy must infer the task identity from collected transitions by modelling its dependency on states, actions and rewards. Because the different datasets may have state-action distributions with large divergence, the task inference module can learn to ignore the rewards and spuriously correlate \textit{only} state-action pairs to the task identity, leading to poor test time performance. To robustify task inference, we propose a novel application of the triplet loss. To mine hard negative examples, we relabel the transitions from the training tasks by approximating their reward functions. When we allow further training on the unseen tasks, using the trained policy as an initialization leads to significantly faster convergence compared to randomly initialized policies (up to 80% improvement and across 5 different Mujoco task distributions). We name our method \textbf{MBML} (\textbf{M}ulti-task \textbf{B}atch RL with \textbf{M}etric \textbf{L}earning).

[On 1/n neural representation and robustness](#)

- Josue Nassar, Piotr Sokol, Sueyeon Chung, Kenneth D. Harris, Il Memming Park
- abstract@[open-review](#): Understanding the nature of representation in neural networks is a goal shared by neuroscience and machine learning. It is therefore exciting that both fields converge not only on shared questions but also on similar approaches. A pressing question in these areas is understanding how the structure of the representation used by neural networks affects both their generalization, and robustness to perturbations. In this work, we investigate the latter by juxtaposing experimental results regarding the covariance spectrum of neural representations in the mouse V1 (Stringer et al) with artificial neural networks. We use adversarial robustness to probe Stringer et alâ€™s theory regarding the causal role of a 1/n covariance spectrum. We empirically investigate the benefits such a neural code confers in neural networks, and illuminate its role in multi-layer architectures. Our results show that imposing the experimentally observed structure on artificial neural networks makes them more robust to adversarial attacks. Moreover, our findings complement the existing theory relating wide neural networks to kernel methods, by showing the role of intermediate representations.

Boundary thickness and robustness in learning models

- Yaoqing Yang, Rajiv Khanna, Yaodong Yu, Amir Gholami, Kurt Keutzer, Joseph E. Gonzalez, Kannan Ramchandran, Michael W. Mahoney
- abstract@[open-review](#): Robustness of machine learning models to various adversarial and non-adversarial corruptions continues to be of interest. In this paper, we introduce the notion of the boundary thickness of a classifier, and we describe its connection with and usefulness for model robustness. Thick decision boundaries lead to improved performance, while thin decision boundaries lead to overfitting (e.g., measured by the robust generalization gap between training and testing) and lower robustness. We show that a thicker boundary helps improve robustness against adversarial examples (e.g., improving the robust test accuracy of adversarial training), as well as so-called out-of-distribution (OOD) transforms, and we show that many commonly-used regularization and data augmentation procedures can increase boundary thickness. On the theoretical side, we establish that maximizing boundary thickness is akin to minimizing the so-called mixup loss. Using these observations, we can show that noise-augmentation on mixup training further increases boundary thickness, thereby combating vulnerability to various forms of adversarial attacks and OOD transforms. We can also show that the performance improvement in several recent lines of work happens in conjunction with a thicker boundary.

Demixed shared component analysis of neural population data from multiple brain areas

- Yu Takagi, Steven Kennerley, Jun-ichiro Hirayama, Laurence Hunt
- abstract@[open-review](#): Recent advances in neuroscience data acquisition allow for the simultaneous recording of large populations of neurons across multiple brain areas while subjects perform complex cognitive tasks. Interpreting these data requires us to index how task-relevant information is shared across brain regions, but this is often confounded by the mixing of different task parameters at the single neuron level. Here, inspired by a method developed for a single brain area, we introduce a new technique for demixing variables across multiple brain areas, called demixed shared component analysis (dSCA). dSCA decomposes population activity into a few components, such that the shared components capture the maximum amount of shared information across brain regions while also depending on relevant task parameters. This yields interpretable components that express which variables are shared between different brain regions and when this information is shared across time. To illustrate our method, we reanalyze two datasets recorded during decision-making tasks in rodents and macaques. We find that dSCA provides new insights into the shared computation between different brain areas in these datasets, relating to several different aspects of decision formation.

Learning Kernel Tests Without Data Splitting

- Jonas Käbler, Wittawat Jitkrittum, Bernhard Schölkopf, Krikamol Muandet
- abstract@[open-review](#): Modern large-scale kernel-based tests such as maximum mean discrepancy (MMD) and kernelized Stein discrepancy (KSD) optimize kernel hyperparameters on a held-out sample via data splitting to obtain the most powerful test statistics. While data splitting results in a tractable null distribution, it suffers from a reduction in test power due to smaller test sample size. Inspired by the selective inference framework, we propose an approach that enables learning the hyperparameters and testing on the full sample without data splitting. Our approach can correctly calibrate the test in the presence of such dependency, and yield a test threshold in closed form. At the same significance level, our approach's test power is empirically larger than that of the data-splitting approach, regardless of its split proportion.

Unsupervised Data Augmentation for Consistency Training

- Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, Quoc Le
- abstract@[open-review](#): Semi-supervised learning lately has shown much promise in improving deep learning models when labeled data is scarce. Common among recent approaches is the use of consistency training on a large amount of unlabeled data to constrain model predictions to be invariant to input noise. In this work, we present a new perspective on how to effectively noise unlabeled examples and argue that the quality of noising, specifically those produced by advanced data augmentation methods, plays a crucial role in semi-supervised learning. By substituting simple noising operations with advanced data augmentation methods such as RandAugment and back-translation, our method brings substantial improvements across six language and three vision tasks under the same consistency training framework. On the IMDb text classification dataset, with only 20 labeled examples, our method achieves an error rate of 4.20, outperforming the state-of-the-art model trained on 25,000 labeled examples. On a standard semi-supervised learning benchmark, CIFAR-10, our method outperforms all previous approaches and achieves an error rate of 5.43 with only 250 examples. Our method also combines well with transfer learning, e.g., when finetuning from BERT, and yields improvements in high-data regime, such as ImageNet, whether when there is only 10% labeled data or when a full labeled set with 1.3M extra unlabeled examples is used. Code is available at <https://github.com/google-research/uda>.

Subgroup-based Rank-1 Lattice Quasi-Monte Carlo

- Yueming LYU, Yuan Yuan, Ivor Tsang
- abstract@[open-review](#): Quasi-Monte Carlo (QMC) is an essential tool for integral approximation, Bayesian inference, and sampling for simulation in science, etc. In the QMC area, the rank-1 lattice is important due to its simple operation, and nice property for point set construction. However, the construction of the generating vector of the rank-1 lattice is usually time-consuming through an exhaustive computer search. To address this issue, we propose a simple closed-form rank-1 lattice construction method based on group theory. Our method reduces the number of distinct pairwise distance values to generate a more regular lattice. We theoretically prove a lower and an upper bound of the minimum pairwise distance of any non-degenerate rank-1 lattice. Empirically, our methods can generate near-optimal rank-1 lattice compared with Korobov exhaustive search regarding the $\| \cdot \|_1$ -norm and $\| \cdot \|_2$ -norm minimum distance. Moreover, experimental results show that our method achieves superior approximation performance on the benchmark integration test problems and the kernel approximation problems.

Minibatch vs Local SGD for Heterogeneous Distributed Learning

- Blake E. Woodworth, Kumar Kshitij Patel, Nati Srebro
- abstract@[open-review](#): We analyze Local SGD (aka parallel or federated SGD) and Minibatch SGD in the heterogeneous distributed setting, where each machine has access to stochastic gradient estimates for a different, machine-specific, convex objective; the goal is to optimize w.r.t.~the average objective; and machines can only communicate intermittently. We argue that, (i) Minibatch SGD (even without acceleration) dominates all existing analysis of Local SGD in this setting, (ii) accelerated Minibatch SGD is optimal when the heterogeneity is high, and (iii) present the first upper bound for Local SGD that improves over Minibatch SGD in a non-homogeneous regime.

Multi-task Causal Learning with Gaussian Processes

- Virginia Aglietti, Theodoros Damoulas, Mauricio Alvarez, Javier González
- abstract@[open-review](#): This paper studies the problem of learning the correlation structure of a set of intervention functions defined on the directed acyclic graph (DAG) of a causal model. This is useful when we are interested in jointly learning the causal effects of interventions on different subsets of variables in a DAG, which is common in field such as healthcare or operations research. We propose the first multi-task causal Gaussian process (GP) model, which we call DAG-GP, that allows for information sharing across continuous interventions and across experiments on different variables. DAG-GP accommodates different assumptions in terms of data availability and captures the correlation between functions lying in input spaces of different dimensionality via a well-defined integral operator. We give theoretical results detailing when and how the DAG-GP model can be formulated depending

on the DAG. We test both the quality of its predictions and its calibrated uncertainties. Compared to single-task models, DAG-GP achieves the best fitting performance in a variety of real and synthetic settings. In addition, it helps to select optimal interventions faster than competing approaches when used within sequential decision making frameworks, like active learning or Bayesian optimization.

[Proximity Operator of the Matrix Perspective Function and its Applications](#)

- Joong-Ho (Johann) Won
- abstract@[open-review](#): We show that the matrix perspective function, which is jointly convex in the Cartesian product of a standard Euclidean vector space and a conformal space of symmetric matrices, has a proximity operator in an almost closed form. The only implicit part is to solve a semismooth, univariate root finding problem. We uncover the connection between our problem of study and the matrix nearness problem. Through this connection, we propose a quadratically convergent Newton algorithm for the root finding problem. Experiments verify that the evaluation of the proximity operator requires at most 8 Newton steps, taking less than 5s for 2000 by 2000 matrices on a standard laptop. Using this routine as a building block, we demonstrate the usefulness of the studied proximity operator in constrained maximum likelihood estimation of Gaussian mean and covariance, pseudolikelihood-based graphical model selection, and a matrix variant of the scaled lasso problem.

[Generative 3D Part Assembly via Dynamic Graph Learning](#)

- jialei huang, Guanqi Zhan, Qingnan Fan, Kaichun Mo, Lin Shao, Baoquan Chen, Leonidas J. Guibas, Hao Dong
- abstract@[open-review](#): Autonomous part assembly is a challenging yet crucial task in 3D computer vision and robotics. Analogous to buying an IKEA furniture, given a set of 3D parts that can assemble a single shape, an intelligent agent needs to perceive the 3D part geometry, reason to propose pose estimations for the input parts, and finally call robotic planning and control routines for actuation. In this paper, we focus on the pose estimation subproblem from the vision side involving geometric and relational reasoning over the input part geometry. Essentially, the task of generative 3D part assembly is to predict a 6-DoF part pose, including a rigid rotation and translation, for each input part that assembles a single 3D shape as the final output. To tackle this problem, we propose an assembly-oriented dynamic graph learning framework that leverages an iterative graph neural network as a backbone. It explicitly conducts sequential part assembly refinements in a coarse-to-fine manner, exploits a pair of part relation reasoning module and part aggregation module for dynamically adjusting both part features and their relations in the part graph. We conduct extensive experiments and quantitative comparisons to three strong baseline methods, demonstrating the effectiveness of the proposed approach.

[Improving Natural Language Processing Tasks with Human Gaze-Guided Neural Attention](#)

- Ekta Sood, Simon Tannert, Philipp Mueller, Andreas Bulling
- abstract@[open-review](#): A lack of corpora has so far limited advances in integrating human gaze data as a supervisory signal in neural attention mechanisms for natural language processing (NLP). We propose a novel hybrid text saliency model (TSM) that, for the first time, combines a cognitive model of reading with explicit human gaze supervision in a single machine learning framework. On four different corpora we demonstrate that our hybrid TSM duration predictions are highly correlated with human gaze ground truth. We further propose a novel joint modeling approach to integrate TSM predictions into the attention layer of a network designed for a specific upstream NLP task without the need for any task-specific human gaze data. We demonstrate that our joint model outperforms the state of the art in paraphrase generation on the Quora Question Pairs corpus by more than 10% in BLEU-4 and achieves state of the art performance for sentence compression on the challenging Google Sentence Compression corpus. As such, our work introduces a practical approach for bridging between data-driven and cognitive models and demonstrates a new way to integrate human gaze-guided neural attention into NLP tasks.

[The Power of Comparisons for Actively Learning Linear Classifiers](#)

- Max Hopkins, Daniel Kane, Shachar Lovett
- abstract@[open-review](#): In the world of big data, large but costly to label datasets dominate many fields. Active learning, a semi-supervised alternative to the standard PAC-learning model, was introduced to explore whether adaptive labeling could learn concepts with exponentially fewer labeled samples. While previous results show that active learning performs no better than its supervised alternative for important concept classes such as linear separators, we show that by adding weak distributional assumptions and allowing comparison queries, active learning requires exponentially fewer samples. Further, we show that these results hold as well for a stronger model of learning called Reliable and Probably Useful (RPU) learning. In this model, our learner is not allowed to make mistakes, but may instead answer ``I don't know.'' While previous negative results showed this model to have intractably large sample complexity for label queries, we show that comparison queries make RPU-learning at worst logarithmically more expensive in both the passive and active regimes.

[From Boltzmann Machines to Neural Networks and Back Again](#)

- Surbhi Goel, Adam Klivans, Frederic Koehler
- abstract@[open-review](#): Graphical models are powerful tools for modeling high-dimensional data, but learning graphical models in the presence of latent variables is well-known to be difficult. In this work we give new results for learning Restricted Boltzmann Machines, probably the most well-studied class of latent variable models. Our results are based on new connections to learning two-layer neural networks under ℓ_{∞} bounded input; for both problems, we give nearly optimal results under the conjectured hardness of sparse parity with noise. Using the connection between RBMs and feedforward networks, we also initiate the theoretical study of em{supervised RBMs} \citet{hinton2012practical}, a version of neural-network learning that couples distributional assumptions induced from the underlying graphical model with the architecture of the unknown function class. We then give an algorithm for learning a natural class of supervised RBMs with better runtime than what is possible for its related class of networks without distributional assumptions.

[Crush Optimism with Pessimism: Structured Bandits Beyond Asymptotic Optimality](#)

- Kwang-Sung Jun, Chicheng Zhang
- abstract@[open-review](#): We study stochastic structured bandits for minimizing regret. The fact that the popular optimistic algorithms do not achieve the asymptotic instance-dependent regret optimality (asymptotic optimality for short) has recently alluded researchers. On the other hand, it is known that one can achieve bounded regret (i.e., does not grow indefinitely with n) in certain instances. Unfortunately, existing asymptotically optimal algorithms rely on forced sampling that introduces an $\Omega(1)$ term w.r.t. the time horizon n in their regret, failing to adapt to the ``easiness'' of the instance. In this paper, we focus on the finite hypothesis case and ask if one can achieve the asymptotic optimality while enjoying bounded regret whenever possible. We provide a positive answer by introducing a new algorithm called CRush Optimism with Pessimism (CROP) that eliminates optimistic hypotheses by pulling the informative arms indicated by a pessimistic hypothesis. Our finite-time analysis shows that CROP (i) achieves a constant-factor asymptotic optimality and, thanks to the forced-exploration-free design, (ii) adapts to bounded regret, and (iii) its regret bound scales not with K but with an effective number of arms K_{ψ} that we introduce. We also discuss a problem class where CROP can be exponentially better than existing algorithms in nonasymptotic regimes. This problem class also reveals a surprising fact that even a clairvoyant oracle who plays according to the asymptotically optimal arm pull scheme may suffer a linear worst-case regret.

[Pruning neural networks without any data by iteratively conserving synaptic flow](#)

- Hidenori Tanaka, Daniel Kunin, Daniel L. Yamins, Surya Ganguli
- abstract@[open-review](#): Pruning the parameters of deep neural networks has generated intense interest due to potential savings in time, memory and energy both during training and at test time. Recent works have identified, through an expensive sequence of training and pruning cycles, the existence of winning lottery tickets or sparse trainable subnetworks at initialization. This raises a foundational question: can we identify highly sparse trainable subnetworks at initialization, without ever training, or indeed without ever looking at the data? We provide an affirmative answer to this question through theory driven algorithm design. We first mathematically formulate and experimentally verify a conservation law that explains why existing gradient-based pruning algorithms at initialization suffer from layer-collapse, the premature pruning of an entire layer rendering a network untrainable. This theory also elucidates how layer-collapse can be entirely avoided, motivating a novel pruning algorithm Iterative Synaptic Flow Pruning (SynFlow). This algorithm can be interpreted as preserving the total flow of synaptic strengths through the network at initialization subject to a sparsity constraint. Notably, this algorithm makes no reference to the training data and consistently competes with or outperforms existing state-of-the-art pruning algorithms at initialization over a range of models (VGG and ResNet), datasets (CIFAR-10/100 and Tiny ImageNet), and sparsity constraints (up to 99.99 percent). Thus our data-agnostic pruning algorithm challenges the existing paradigm that, at initialization, data must be used to quantify which synapses are important.

[Detecting Interactions from Neural Networks via Topological Analysis](#)

- Zirui Liu, Qingquan Song, Kaixiong Zhou, Ting-Hsiang Wang, Ying Shan, Xia Hu
- abstract@[open-review](#): Detecting statistical interactions between input features is a crucial and challenging task. Recent advances demonstrate that it is possible to extract learned interactions from trained neural networks. It has also been observed that, in neural networks, any interacting features must follow a strongly weighted connection to common hidden units. Motivated by the observation, in this paper, we propose to investigate the interaction detection problem from a novel topological perspective by analyzing the connectivity in neural networks. Specially, we propose a new measure for quantifying interaction strength, based upon the well-received theory of persistent homology. Based on this measure, a Persistence Interaction Detection (PID) algorithm is developed to efficiently detect interactions. Our proposed algorithm is evaluated across a number of interaction detection tasks on several synthetic and real-world datasets with different hyperparameters. Experimental results validate that the PID algorithm outperforms the state-of-the-art baselines.

[Neural Bridge Sampling for Evaluating Safety-Critical Autonomous Systems](#)

- Aman Sinha, Matthew O'Kelly, Russ Tedrake, John C. Duchi
- abstract@[open-review](#): Learning-based methodologies increasingly find applications in safety-critical domains like autonomous driving and medical robotics. Due to the rare nature of dangerous events, real-world testing is prohibitively expensive and unscalable. In this work, we employ a probabilistic approach to safety evaluation in simulation, where we are concerned with computing the probability of dangerous events. We develop a novel rare-event simulation method that combines exploration, exploitation, and optimization techniques to find failure modes and estimate their rate of occurrence. We provide rigorous guarantees for the performance of our method in terms of both statistical and computational efficiency. Finally, we demonstrate the efficacy of our approach on a variety of scenarios, illustrating its usefulness as a tool for rapid sensitivity analysis and model comparison that are essential to developing and testing safety-critical autonomous systems.

[Interpretable and Personalized Apprenticeship Scheduling: Learning Interpretable Scheduling Policies from Heterogeneous User Demonstrations](#)

- Rohan Paleja, Andrew Silva, Letian Chen, Matthew Gombolay
- abstract@[open-review](#): Resource scheduling and coordination is an NP-hard optimization requiring an efficient allocation of agents to a set of tasks with upper- and lower bound temporal and resource constraints. Due to the large-scale and dynamic nature of resource coordination in hospitals and factories, human domain experts manually plan and adjust schedules on the fly. To perform this job, domain experts leverage heterogeneous strategies and rules-of-thumb honed over years of apprenticeship. What is critically needed is the ability to extract this domain knowledge in a heterogeneous and interpretable apprenticeship learning framework to scale beyond the power of a single human expert, a necessity in safety-critical domains. We propose a personalized and interpretable apprenticeship scheduling algorithm that infers an interpretable representation of all human task demonstrators by extracting decision-making criteria via an inferred, personalized embedding non-parametric in the number of demonstrator types. We achieve near-perfect LfD accuracy in synthetic domains and 88.22% accuracy on a planning domain with real-world data, outperforming baselines. Finally, our user study showed our methodology produces more interpretable and easier-to-use models than neural networks ($p < 0.05$).

[Task-Agnostic Online Reinforcement Learning with an Infinite Mixture of Gaussian Processes](#)

- Mengdi Xu, Wenhao Ding, Jiacheng Zhu, ZUXIN LIU, Baiming Chen, Ding Zhao
- abstract@[open-review](#): Continuously learning to solve unseen tasks with limited experience has been extensively pursued in meta-learning and continual learning, but with restricted assumptions such as accessible task distributions, independently and identically distributed tasks, and clear task delineations. However, real-world physical tasks frequently violate these assumptions, resulting in performance degradation. This paper proposes a continual online model-based reinforcement learning approach that does not require pre-training to solve task-agnostic problems with unknown task boundaries. We maintain a mixture of experts to handle nonstationarity, and represent each different type of dynamics with a Gaussian Process to efficiently leverage collected data and expressively model uncertainty. We propose a transition prior to account for the temporal dependencies in streaming data and update the mixture online via sequential variational inference. Our approach reliably handles the task distribution shift by generating new models for never-before-seen dynamics and reusing old models for previously seen dynamics. In experiments, our approach outperforms alternative methods in non-stationary tasks, including classic control with changing dynamics and decision making in different driving scenarios.

[Benchmarking Deep Learning Interpretability in Time Series Predictions](#)

- Aya Abdelsalam Ismail, Mohamed Gunady, Hector Corrada Bravo, Soheil Feizi
- abstract@[open-review](#): Saliency methods are used extensively to highlight the importance of input features in model predictions. These methods are mostly used in vision and language tasks, and their applications to time series data is relatively unexplored. In this paper, we set out to extensively compare the performance of various saliency-based interpretability methods across diverse neural architectures, including Recurrent Neural Network, Temporal Convolutional Networks, and Transformers in a new benchmark of synthetic time series data. We propose and report multiple metrics to empirically evaluate the performance of saliency methods for detecting feature importance over time using both precision (i.e., whether identified features contain meaningful signals) and recall (i.e., the number of features with signal identified as important). Through several experiments, we show that (i) in general, network architectures and saliency methods fail to reliably and accurately identify feature importance over time in time series data, (ii) this failure is mainly due to the conflation of time and feature domains, and (iii) the quality of saliency maps can be improved substantially by using our proposed two-step temporal saliency rescaling (TSR) approach that first calculates the importance of each time step before calculating the importance of each feature at a time step.

[Federated Principal Component Analysis](#)

- Andreas Grammenos, Rodrigo Mendoza Smith, Jon Crowcroft, Cecilia Mascolo
- abstract@[open-review](#): We present a federated, asynchronous, and $(\backslash \epsilon, \delta)$ -differentially private algorithm for \$PCA\$ in the memory-limited setting. % Our algorithm incrementally computes local model updates using a streaming procedure and adaptively estimates its \$r\$ leading

principal components when only $\mathcal{O}(dr)$ memory is available with d being the dimensionality of the data. % We guarantee differential privacy via an input-perturbation scheme in which the covariance matrix of a dataset $B\{X\} \in \mathbb{R}^{d \times n}$ is perturbed with a non-symmetric random Gaussian matrix with variance in $\mathcal{O}(\left(\frac{d}{n}\right)^2 \log d)$, thus improving upon the state-of-the-art. % Furthermore, contrary to previous federated or distributed algorithms for PCA , our algorithm is also invariant to permutations in the incoming data, which provides robustness against straggler or failed nodes. % Numerical simulations show that, while using limited-memory, our algorithm exhibits performance that closely matches or outperforms traditional non-federated algorithms, and in the absence of communication latency, it exhibits attractive horizontal scalability.

[\(De\)Randomized Smoothing for Certifiable Defense against Patch Attacks](#)

- Alexander Levine, Soheil Feizi
- abstract@[open-review](#): Patch adversarial attacks on images, in which the attacker can distort pixels within a region of bounded size, are an important threat model since they provide a quantitative model for physical adversarial attacks. In this paper, we introduce a certifiable defense against patch attacks that guarantees for a given image and patch attack size, no patch adversarial examples exist. Our method is related to the broad class of randomized smoothing robustness schemes which provide high-confidence probabilistic robustness certificates. By exploiting the fact that patch attacks are more constrained than general sparse attacks, we derive meaningfully large robustness certificates against them. Additionally, in contrast to smoothing-based defenses against L_p and sparse attacks, our defense method against patch attacks is de-randomized, yielding improved, deterministic certificates. Compared to the existing patch certification method proposed by (Chiang et al., 2020), which relies on interval bound propagation, our method can be trained significantly faster, achieves high clean and certified robust accuracy on CIFAR-10, and provides certificates at ImageNet scale. For example, for a 5-by-5 patch attack on CIFAR-10, our method achieves up to around 57.6% certified accuracy (with a classifier with around 83.8% clean accuracy), compared to at most 30.3% certified accuracy for the existing method (with a classifier with around 47.8% clean accuracy). Our results effectively establish a new state-of-the-art of certifiable defense against patch attacks on CIFAR-10 and ImageNet.

[SMYRF - Efficient Attention using Asymmetric Clustering](#)

- Giannis Daras, Nikita Kitaev, Augustus Odena, Alexandros G. Dimakis
- abstract@[open-review](#): We propose a novel type of balanced clustering algorithm to approximate attention. Attention complexity is reduced from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \log N)$, where N is the sequence length. Our algorithm, SMYRF, uses Locality Sensitive Hashing (LSH) in a novel way by defining new Asymmetric transformations and an adaptive scheme that produces balanced clusters. The biggest advantage of SMYRF is that it can be used as a drop-in replacement for dense attention layers without any retraining. On the contrary, prior fast attention methods impose constraints (e.g. tight queries and keys) and require re-training from scratch. We apply our method to pre-trained state-of-the-art Natural Language Processing and Computer Vision models and we report significant memory and speed benefits. Notably, SMYRF-BERT outperforms (slightly) BERT on GLUE, while using 50% less memory. We also show that SMYRF can be used interchangeably with dense attention before and after training. Finally, we use SMYRF to train GANs with attention in high resolutions. Using a single TPU, we train BigGAN on Celeba-HQ, with attention at resolution 128x128 and 256x256, capable of generating realistic human faces.

[Introducing Routing Uncertainty in Capsule Networks](#)

- Fabio De Sousa Ribeiro, Georgios Leontidis, Stefanos Kollias
- abstract@[open-review](#): Rather than performing inefficient local iterative routing between adjacent capsule layers, we propose an alternative global view based on representing the inherent uncertainty in part-object assignment. In our formulation, the local routing iterations are replaced with variational inference of part-object connections in a probabilistic capsule network, leading to a significant speedup without sacrificing performance. In this way, global context is also considered when routing capsules by introducing global latent variables that have direct influence on the objective function, and are updated discriminatively in accordance with the minimum description length (MDL) principle. We focus on enhancing capsule network properties, and perform a thorough evaluation on pose-aware tasks, observing improvements in performance over previous approaches whilst being more computationally efficient.

[A Simple and Efficient Smoothing Method for Faster Optimization and Local Exploration](#)

- Kevin Scaman, Ludovic DOS SANTOS, Merwan Barlier, Igor Colin
- abstract@[open-review](#): This work proposes a novel smoothing method, called Bend, Mix and Release (BMR), that extends two well-known smooth approximations of the convex optimization literature: randomized smoothing and the Moreau envelope. The BMR smoothing method allows to trade-off between the computational simplicity of randomized smoothing (RS) and the approximation efficiency of the Moreau envelope (ME). More specifically, we show that BMR achieves up to a \sqrt{d} multiplicative improvement compared to the approximation error of RS, where d is the dimension of the search space, while being less computation intensive than the ME. For non-convex objectives, BMR also has the desirable property to widen local minima, allowing optimization methods to reach small cracks and crevices of extremely irregular and non-convex functions, while being well-suited to a distributed setting. This novel smoothing method is then used to improve first-order non-smooth optimization (both convex and non-convex) by allowing for a local exploration of the search space. More specifically, our analysis sheds light on the similarities between evolution strategies and BMR, creating a link between exploration strategies of zeroth-order methods and the regularity of first-order optimization problems. Finally, we evidence the impact of BMR through synthetic experiments.

[Hyperparameter Ensembles for Robustness and Uncertainty Quantification](#)

- Florian Wenzel, Jasper Snoek, Dustin Tran, Rodolphe Jenatton
- abstract@[open-review](#): Ensembles over neural network weights trained from different random initialization, known as deep ensembles, achieve state-of-the-art accuracy and calibration. The recently introduced batch ensembles provide a drop-in replacement that is more parameter efficient. In this paper, we design ensembles not only over weights, but over hyperparameters to improve the state of the art in both settings. For best performance independent of budget, we propose hyper-deep ensembles, a simple procedure that involves a random search over different hyperparameters, themselves stratified across multiple random initializations. Its strong performance highlights the benefit of combining models with both weight and hyperparameter diversity. We further propose a parameter efficient version, hyper-batch ensembles, which builds on the layer structure of batch ensembles and self-tuning networks. The computational and memory costs of our method are notably lower than typical ensembles. On image classification tasks, with MLP, LeNet, ResNet 20 and Wide ResNet 28-10 architectures, we improve upon both deep and batch ensembles.

[Neutralizing Self-Selection Bias in Sampling for Sortition](#)

- Bailey Flanigan, Paul Gál, Anupam Gupta, Ariel D. Procaccia
- abstract@[open-review](#): Sortition is a political system in which decisions are made by panels of randomly selected citizens. The process for selecting a sortition panel is traditionally thought of as uniform sampling without replacement, which has strong fairness properties. In practice, however, sampling without replacement is not possible since only a fraction of agents is willing to participate in a panel when invited, and different demographic groups participate at different rates. In order to still produce panels whose composition resembles that of the population, we develop a sampling algorithm that restores close-to-equal representation probabilities for all agents while satisfying meaningful demographic quotas. As part of its input, our algorithm requires probabilities indicating how likely each volunteer in the pool was to participate. Since these participation probabilities are not directly observable,

we show how to learn them, and demonstrate our approach using data on a real sortition panel combined with information on the general population in the form of publicly available survey data.

[On the Convergence of Smooth Regularized Approximate Value Iteration Schemes](#)

- Elena Smirnova, Elvis Dohmatob
- abstract@[open-review](#): Entropy regularization, smoothing of Q-values and neural network function approximator are key components of the state-of-the-art reinforcement learning (RL) algorithms, such as Soft Actor-Critic~[cite{haarnoja2018soft}](#). Despite the widespread use, the impact of these core techniques on the convergence of RL algorithms is not yet fully understood. In this work, we analyse these techniques from error propagation perspective using the approximate dynamic programming framework. In particular, our analysis shows that (1) value smoothing results in increased stability of the algorithm in exchange for slower convergence, (2) entropy regularization reduces overestimation errors at the cost of modifying the original problem, (3) we study a combination of these techniques that describes the Soft Actor-Critic algorithm.

[Off-Policy Evaluation via the Regularized Lagrangian](#)

- Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, Dale Schuurmans
- abstract@[open-review](#): The recently proposed distribution correction estimation (DICE) family of estimators has advanced the state of the art in off-policy evaluation from behavior-agnostic data. While these estimators all perform some form of stationary distribution correction, they arise from different derivations and objective functions. In this paper, we unify these estimators as regularized Lagrangians of the same linear program. The unification allows us to expand the space of DICE estimators to new alternatives that demonstrate improved performance. More importantly, by analyzing the expanded space of estimators both mathematically and empirically we find that dual solutions offer greater flexibility in navigating the tradeoff between optimization stability and estimation bias, and generally provide superior estimates in practice.

[The LoCA Regret: A Consistent Metric to Evaluate Model-Based Behavior in Reinforcement Learning](#)

- Harm Van Seijen, Hadi Nekoei, Evan Racah, Sarath Chandar
- abstract@[open-review](#): Deep model-based Reinforcement Learning (RL) has the potential to substantially improve the sample-efficiency of deep RL. While various challenges have long held it back, a number of papers have recently come out reporting success with deep model-based methods. This is a great development, but the lack of a consistent metric to evaluate such methods makes it difficult to compare various approaches. For example, the common single-task sample-efficiency metric conflates improvements due to model-based learning with various other aspects, such as representation learning, making it difficult to assess true progress on model-based RL. To address this, we introduce an experimental setup to evaluate model-based behavior of RL methods, inspired by work from neuroscience on detecting model-based behavior in humans and animals. Our metric based on this setup, the Local Change Adaptation (LoCA) regret, measures how quickly an RL method adapts to a local change in the environment. Our metric can identify model-based behavior, even if the method uses a poor representation and provides insight in how close a method's behavior is from optimal model-based behavior. We use our setup to evaluate the model-based behavior of MuZero on a variation of the classic Mountain Car task.

[Neural Power Units](#)

- Niklas Heim, Tomas Pevny, Vasek Smidl
- abstract@[open-review](#): Conventional Neural Networks can approximate simple arithmetic operations, but fail to generalize beyond the range of numbers that were seen during training. Neural Arithmetic Units aim to overcome this difficulty, but current arithmetic units are either limited to operate on positive numbers or can only represent a subset of arithmetic operations. We introduce the Neural Power Unit (NPU) that operates on the full domain of real numbers and is capable of learning arbitrary power functions in a single layer. The NPU thus fixes the shortcomings of existing arithmetic units and extends their expressivity. We achieve this by using complex arithmetic without requiring a conversion of the network to complex numbers. A simplification of the unit to the RealNPU yields a highly transparent model. We show that the NPUs outperform their competitors in terms of accuracy and sparsity on artificial arithmetic datasets, and that the RealNPU can discover the governing equations of a dynamical system only from data.

[Towards Scalable Bayesian Learning of Causal DAGs](#)

- Jussi Viinikka, Antti Hyttinen, Johan Pensar, Mikko Koivisto
- abstract@[open-review](#): We give methods for Bayesian inference of directed acyclic graphs, DAGs, and the induced causal effects from passively observed complete data. Our methods build on a recent Markov chain Monte Carlo scheme for learning Bayesian networks, which enables efficient approximate sampling from the graph posterior, provided that each node is assigned a small number K of candidate parents. We present algorithmic techniques to significantly reduce the space and time requirements, which make the use of substantially larger values of K feasible. Furthermore, we investigate the problem of selecting the candidate parents per node so as to maximize the covered posterior mass. Finally, we combine our sampling method with a novel Bayesian approach for estimating causal effects in linear Gaussian DAG models. Numerical experiments demonstrate the performance of our methods in detecting ancestor-descendant relations, and in causal effect estimation our Bayesian method is shown to outperform previous approaches.

[A Dictionary Approach to Domain-Invariant Learning in Deep Networks](#)

- Ze Wang, Xiuyuan Cheng, Guillermo Sapiro, Qiang Qiu
- abstract@[open-review](#): In this paper, we consider domain-invariant deep learning by explicitly modeling domain shifts with only a small amount of domain-specific parameters in a Convolutional Neural Network (CNN). By exploiting the observation that a convolutional filter can be well approximated as a linear combination of a small set of dictionary atoms, we show for the first time, both empirically and theoretically, that domain shifts can be effectively handled by decomposing a convolutional layer into a domain-specific atom layer and a domain-shared coefficient layer, while both remain convolutional. An input channel will now first convolve spatially only with each respective domain-specific dictionary atom to ``absorb'' domain variations, and then output channels are linearly combined using common decomposition coefficients trained to promote shared semantics across domains. We use toy examples, rigorous analysis, and real-world examples with diverse datasets and architectures, to show the proposed plug-in framework's effectiveness in cross and joint domain performance and domain adaptation. With the proposed architecture, we need only a small set of dictionary atoms to model each additional domain, which brings a negligible amount of additional parameters, typically a few hundred.

[Bootstrapping neural processes](#)

- Juho Lee, Yoonho Lee, Jungtaek Kim, Eunho Yang, Sung Ju Hwang, Yee Whye Teh
- abstract@[open-review](#): Unlike in the traditional statistical modeling for which a user typically hand-specify a prior, Neural Processes (NPs) implicitly define a broad class of stochastic processes with neural networks. Given a data stream, NP learns a stochastic process that best describes the data. While this ``data-driven'' way of learning stochastic processes has proven to handle various types of data, NPs still relies on an assumption that uncertainty in stochastic processes is modeled by a single latent variable, which potentially limits the flexibility. To this end, we propose the Bootstrapping Neural Process (BNP), a novel extension of the NP family using the bootstrap. The bootstrap is a classical data-driven technique for estimating uncertainty, which allows BNP to learn the stochasticity in NPs without assuming a particular form. We demonstrate the efficacy of BNP on various types of data and its robustness in the presence of model-data mismatch.

[Large-Scale Adversarial Training for Vision-and-Language Representation Learning](#)

- Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, Jingjing Liu
- abstract@[open-review](#): We present VILLA, the first known effort on large-scale adversarial training for vision-and-language (V+L) representation learning. VILLA consists of two training stages: (i) task-agnostic adversarial pre-training; followed by (ii) task-specific adversarial finetuning. Instead of adding adversarial perturbations on image pixels and textual tokens, we propose to perform adversarial training in the embedding space of each modality. To enable large-scale training, we adopt the ``free'' adversarial training strategy, and combine it with KL-divergence-based regularization to promote higher invariance in the embedding space. We apply VILLA to current best-performing V+L models, and achieve new state of the art on a wide range of tasks, including Visual Question Answering, Visual Commonsense Reasoning, Image-Text Retrieval, Referring Expression Comprehension, Visual Entailment, and NLVR2.

[Most ReLU Networks Suffer from \\$ell^2\\$ Adversarial Perturbations](#)

- Amit Daniely, Hadas Shacham
- abstract@[open-review](#): We consider ReLU networks with random weights, in which the dimension decreases at each layer. We show that for most such networks, most examples \mathbf{x} admit an adversarial perturbation at an Euclidean distance of $O(\left(\frac{\|\mathbf{x}\|}{\sqrt{d}}\right)^2)$, where d is the input dimension. Moreover, this perturbation can be found via gradient flow, as well as gradient descent with sufficiently small steps. This result can be seen as an explanation to the abundance of adversarial examples, and to the fact that they are found via gradient descent.

[Compositional Visual Generation with Energy Based Models](#)

- Yilun Du, Shuang Li, Igor Mordatch
- abstract@[open-review](#): A vital aspect of human intelligence is the ability to compose increasingly complex concepts out of simpler ideas, enabling both rapid learning and adaptation of knowledge. In this paper we show that energy-based models can exhibit this ability by directly combining probability distributions. Samples from the combined distribution correspond to compositions of concepts. For example, given a distribution for smiling faces, and another for male faces, we can combine them to generate smiling male faces. This allows us to generate natural images that simultaneously satisfy conjunctions, disjunctions, and negations of concepts. We evaluate compositional generation abilities of our model on the CelebA dataset of natural faces and synthetic 3D scene images. We also demonstrate other unique advantages of our model, such as the ability to continually learn and incorporate new concepts, or infer compositions of concept properties underlying an image.

[Factor Graph Grammars](#)

- David Chiang, Darcey Riley
- abstract@[open-review](#): We propose the use of hyperedge replacement graph grammars for factor graphs, or factor graph grammars (FGGs) for short. FGGs generate sets of factor graphs and can describe a more general class of models than plate notation, dynamic graphical models, case-factor diagrams, and sum-product networks can. Moreover, inference can be done on FGGs without enumerating all the generated factor graphs. For finite variable domains (but possibly infinite sets of graphs), a generalization of variable elimination to FGGs allows exact and tractable inference in many situations. For finite sets of graphs (but possibly infinite variable domains), a FGG can be converted to a single factor graph amenable to standard inference techniques.

[Erdos Goes Neural: an Unsupervised Learning Framework for Combinatorial Optimization on Graphs](#)

- Nikolaos Karalias, Andreas Loukas
- abstract@[open-review](#): Combinatorial optimization (CO) problems are notoriously challenging for neural networks, especially in the absence of labeled instances. This work proposes an unsupervised learning framework for CO problems on graphs that can provide integral solutions of certified quality. Inspired by Erdos' probabilistic method, we use a neural network to parametrize a probability distribution over sets. Crucially, we show that when the network is optimized w.r.t. a suitably chosen loss, the learned distribution contains, with controlled probability, a low-cost integral solution that obeys the constraints of the combinatorial problem. The probabilistic proof of existence is then derandomized to decode the desired solutions. We demonstrate the efficacy of this approach to obtain valid solutions to the maximum clique problem and to perform local graph clustering. Our method achieves competitive results on both real datasets and synthetic hard instances.

[Autoregressive Score Matching](#)

- Chenlin Meng, Lantao Yu, Yang Song, Jiaming Song, Stefano Ermon
- abstract@[open-review](#): Autoregressive models use chain rule to define a joint probability distribution as a product of conditionals. These conditionals need to be normalized, imposing constraints on the functional families that can be used. To increase flexibility, we propose autoregressive conditional score models (AR-CSM) where we parameterize the joint distribution in terms of the derivatives of univariate log-conditionals (scores), which need not be normalized. To train AR-CSM, we introduce a new divergence between distributions named Composite Score Matching (CSM). For AR-CSM models, this divergence between data and model distributions can be computed and optimized efficiently, requiring no expensive sampling or adversarial training. Compared to previous score matching algorithms, our method is more scalable to high dimensional data and more stable to optimize. We show with extensive experimental results that it can be applied to density estimation on synthetic data, image generation, image denoising, and training latent variable models with implicit encoders.

[Debiasing Distributed Second Order Optimization with Surrogate Sketching and Scaled Regularization](#)

- Michal Derezhinski, Burak Bartan, Mert Pilanci, Michael W. Mahoney
- abstract@[open-review](#): In distributed second order optimization, a standard strategy is to average many local estimates, each of which is based on a small sketch or batch of the data. However, the local estimates on each machine are typically biased, relative to the full solution on all of the data, and this can limit the effectiveness of averaging.

Here, we introduce a new technique for debiasing the local estimates, which leads to both theoretical and empirical improvements in the convergence rate of distributed second order methods. Our technique has two novel components: (1) modifying standard sketching techniques to obtain what we call a surrogate sketch; and (2) carefully scaling the global regularization parameter for local computations. Our surrogate sketches are based on determinantal point processes, a family of distributions for which the bias of an estimate of the inverse Hessian can be computed exactly. Based on this computation, we show that when the objective being minimized is l_2 -regularized with parameter λ and individual machines are each given a sketch of size m , then to eliminate the bias, local estimates should be computed using a shrunk regularization parameter given by $\lambda' = \lambda(1 - \frac{d_\lambda}{m})$, where d_λ is the λ -effective dimension of the Hessian (or, for quadratic problems, the data matrix).

[Neural Controlled Differential Equations for Irregular Time Series](#)

- Patrick Kidger, James Morrill, James Foster, Terry Lyons
- abstract@[open-review](#): Neural ordinary differential equations are an attractive option for modelling temporal dynamics. However, a fundamental issue is that the solution to an ordinary differential equation is determined by its initial condition, and there is no mechanism for adjusting the trajectory based on

subsequent observations. Here, we demonstrate how this may be resolved through the well-understood mathematics of \emph{controlled differential equations}. The resulting \emph{neural controlled differential equation} model is directly applicable to the general setting of partially-observed irregularly-sampled multivariate time series, and (unlike previous work on this problem) it may utilise memory-efficient adjoint-based backpropagation even across observations. We demonstrate that our model achieves state-of-the-art performance against similar (ODE or RNN based) models in empirical studies on a range of datasets. Finally we provide theoretical results demonstrating universal approximation, and that our model subsumes alternative ODE models.

[On Efficiency in Hierarchical Reinforcement Learning](#)

- Zheng Wen, Doina Precup, Morteza Ibrahimi, Andre Barreto, Benjamin Van Roy, Satinder Singh
- abstract@[open-review](#): Hierarchical Reinforcement Learning (HRL) approaches promise to provide more efficient solutions to sequential decision making problems, both in terms of statistical as well as computational efficiency. While this has been demonstrated empirically over time in a variety of tasks, theoretical results quantifying the benefits of such methods are still few and far between. In this paper, we discuss the kind of structure in a Markov decision process which gives rise to efficient HRL methods. Specifically, we formalize the intuition that HRL can exploit well repeating "subMDPs", with similar reward and transition structure. We show that, under reasonable assumptions, a model-based Thompson sampling-style HRL algorithm that exploits this structure is statistically efficient, as established through a finite-time regret bound. We also establish conditions under which planning with structure-induced options is near-optimal and computationally efficient.

[On Correctness of Automatic Differentiation for Non-Differentiable Functions](#)

- Wonyeol Lee, Hangyeol Yu, Xavier Rival, Hongseok Yang
- abstract@[open-review](#): Differentiation lies at the core of many machine-learning algorithms, and is well-supported by popular autodiff systems, such as TensorFlow and PyTorch. Originally, these systems have been developed to compute derivatives of differentiable functions, but in practice, they are commonly applied to functions with non-differentiabilities. For instance, neural networks using ReLU define non-differentiable functions in general, but the gradients of losses involving those functions are computed using autodiff systems in practice. This status quo raises a natural question: are autodiff systems correct in any formal sense when they are applied to such non-differentiable functions? In this paper, we provide a positive answer to this question. Using counterexamples, we first point out flaws in often-used informal arguments, such as: non-differentiabilities arising in deep learning do not cause any issues because they form a measure-zero set. We then investigate a class of functions, called PAP functions, that includes nearly all (possibly non-differentiable) functions in deep learning nowadays. For these PAP functions, we propose a new type of derivatives, called intensional derivatives, and prove that these derivatives always exist and coincide with standard derivatives for almost all inputs. We also show that these intensional derivatives are what most autodiff systems compute or try to compute essentially. In this way, we formally establish the correctness of autodiff systems applied to non-differentiable functions.

[Probabilistic Linear Solvers for Machine Learning](#)

- Jonathan Wenger, Philipp Hennig
- abstract@[open-review](#): Linear systems are the bedrock of virtually all numerical computation. Machine learning poses specific challenges for the solution of such systems due to their scale, characteristic structure, stochasticity and the central role of uncertainty in the field. Unifying earlier work we propose a class of probabilistic linear solvers which jointly infer the matrix, its inverse and the solution from matrix-vector product observations. This class emerges from a fundamental set of desiderata which constrains the space of possible algorithms and recovers the method of conjugate gradients under certain conditions. We demonstrate how to incorporate prior spectral information in order to calibrate uncertainty and experimentally showcase the potential of such solvers for machine learning.

[Dynamic Regret of Policy Optimization in Non-Stationary Environments](#)

- Yingjie Fei, Zhuoran Yang, Zhaoran Wang, Qiaomin Xie
- abstract@[open-review](#): We consider reinforcement learning (RL) in episodic MDPs with adversarial full-information reward feedback and unknown fixed transition kernels. We propose two model-free policy optimization algorithms, POWER and POWER++, and establish guarantees for their dynamic regret. Compared with the classical notion of static regret, dynamic regret is a stronger notion as it explicitly accounts for the non-stationarity of environments. The dynamic regret attained by the proposed algorithms interpolates between different regimes of non-stationarity, and moreover satisfies a notion of adaptive (near-)optimality, in the sense that it matches the (near-)optimal static regret under slow-changing environments. The dynamic regret bound features two components, one arising from exploration, which deals with the uncertainty of transition kernels, and the other arising from adaptation, which deals with non-stationary environments. Specifically, we show that POWER++ improves over POWER on the second component of the dynamic regret by actively adapting to non-stationarity through prediction. To the best of our knowledge, our work is the first dynamic regret analysis of model-free RL algorithms in non-stationary environments.

[Multipole Graph Neural Operator for Parametric Partial Differential Equations](#)

- Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Andrew Stuart, Kaushik Bhattacharya, Anima Anandkumar
- abstract@[open-review](#): One of the main challenges in using deep learning-based methods for simulating physical systems and solving partial differential equations (PDEs) is formulating physics-based data in the desired structure for neural networks. Graph neural networks (GNNs) have gained popularity in this area since graphs offer a natural way of modeling particle interactions and provide a clear way of discretizing the continuum models. However, the graphs constructed for approximating such tasks usually ignore long-range interactions due to unfavorable scaling of the computational complexity with respect to the number of nodes. The errors due to these approximations scale with the discretization of the system, thereby not allowing for generalization under mesh-refinement. Inspired by the classical multipole methods, we propose a novel multi-level graph neural network framework that captures interaction at all ranges with only linear complexity. Our multi-level formulation is equivalent to recursively adding inducing points to the kernel matrix, unifying GNNs with multi-resolution matrix factorization of the kernel. Experiments confirm our multi-graph network learns discretization-invariant solution operators to PDEs and can be evaluated in linear time.

[BlockGAN: Learning 3D Object-aware Scene Representations from Unlabelled Images](#)

- Thu H. Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, Niloy Mitra
- abstract@[open-review](#): We present BlockGAN, an image generative model that learns object-aware 3D scene representations directly from unlabelled 2D images. Current work on scene representation learning either ignores scene background or treats the whole scene as one object. Meanwhile, work that considers scene compositionality treats scene objects only as image patches or 2D layers with alpha maps. Inspired by the computer graphics pipeline, we design BlockGAN to learn to first generate 3D features of background and foreground objects, then combine them into 3D features for the whole scene, and finally render them into realistic images. This allows BlockGAN to reason over occlusion and interaction between objects' appearance, such as shadow and lighting, and provides control over each object's 3D pose and identity, while maintaining image realism. BlockGAN is trained end-to-end, using only unlabelled single images, without the need for 3D geometry, pose labels, object masks, or multiple views of the same scene. Our experiments show that using explicit 3D features to represent objects allows BlockGAN to learn disentangled representations both in terms of objects (foreground and background) and their properties (pose and identity).

Online Structured Meta-learning

- Huaxiu Yao, Yingbo Zhou, Mehrdad Mahdavi, Zhenhui (Jessie) Li, Richard Socher, Caiming Xiong
- abstract@[open-review](#): Learning quickly is of great importance for machine intelligence deployed in online platforms. With the capability of transferring knowledge from learned tasks, meta-learning has shown its effectiveness in online scenarios by continuously updating the model with the learned prior. However, current online meta-learning algorithms are limited to learn a globally-shared meta-learner, which may lead to sub-optimal results when the tasks contain heterogeneous information that are difficult to share. We overcome this limitation by proposing an online structured meta-learning (OSML) framework. Inspired by the knowledge organization of human and hierarchical feature representation, OSML explicitly disentangles the meta-learner as a meta-hierarchical graph with different knowledge blocks. When a new task is encountered, it constructs a meta-knowledge pathway by either utilizing the most relevant knowledge blocks or exploring new blocks. Through the meta-knowledge pathway, the model is able to quickly adapt to the new task. In addition, new knowledge is further incorporated into the selected blocks. Experiments on three datasets empirically demonstrate the effectiveness and interpretability of our proposed framework, not only under heterogeneous tasks but also under homogeneous settings.

Learning Strategic Network Emergence Games

- Rakshit Trivedi, Hongyuan Zha
- abstract@[open-review](#): Real-world networks, especially the ones that emerge due to actions of animate agents (e.g. humans, animals), are the result of underlying strategic mechanisms aimed at maximizing individual or collective benefits. Learning approaches built to capture these strategic insights would gain interpretability and flexibility benefits that are required to generalize beyond observations. To this end, we consider a game-theoretic formalism of network emergence that accounts for the underlying strategic mechanisms and take it to the observed data. We propose MINE (Multi-agent Inverse models of Network Emergence mechanism), a new learning framework that solves Markov-Perfect network emergence games using multi-agent inverse reinforcement learning. MINE jointly discovers agents' strategy profiles in the form of network emergence policy and the latent payoff mechanism in the form of learned reward function. In the experiments, we demonstrate that MINE learns versatile payoff mechanisms that: highly correlates with the ground truth for a synthetic case; can be used to analyze the observed network structure; and enable effective transfer in specific settings. Further, we show that the network emergence game as a learned model supports meaningful strategic predictions, thereby signifying its applicability to a variety of network analysis tasks.

Towards Interpretable Natural Language Understanding with Explanations as Latent Variables

- Wangchunshu Zhou, Jinyi Hu, Hanlin Zhang, Xiaodan Liang, Maosong Sun, Chenyan Xiong, Jian Tang
- abstract@[open-review](#): Recently generating natural language explanations has shown very promising results in not only offering interpretable explanations but also providing additional information and supervision for prediction. However, existing approaches usually require a large set of human annotated explanations for training while collecting a large set of explanations is not only time consuming but also expensive. In this paper, we develop a general framework for interpretable natural language understanding that requires only a small set of human annotated explanations for training. Our framework treats natural language explanations as latent variables that model the underlying reasoning process of a neural model. We develop a variational EM framework for optimization where an explanation generation module and an explanation-augmented prediction module are alternatively optimized and mutually enhance each other. Moreover, we further propose an explanation-based self-training method under this framework for semi-supervised learning. It alternates between assigning pseudo-labels to unlabeled data and generating new explanations to iteratively improve each other. Experiments on two natural language understanding tasks demonstrate that our framework can not only make effective predictions in both supervised and semi-supervised settings, but is also able to generate good natural language explanations.

The Mean-Squared Error of Double Q-Learning

- Wentao Weng, Harsh Gupta, Niao He, Lei Ying, R. Srikant
- abstract@[open-review](#): In this paper, we establish a theoretical comparison between the asymptotic mean square errors of double Q-learning and Q-learning. Our result builds upon an analysis for linear stochastic approximation based on Lyapunov equations and applies to both tabular setting or with linear function approximation, provided that the optimal policy is unique and the algorithms converge. We show that the asymptotic mean-square error of Double Q-learning is exactly equal to that of Q-learning if Double Q-learning uses twice the learning rate of Q-learning and the output of Double Q-learning is the average of its two estimators. We also present some practical implications of this theoretical observation using simulations.

What Makes for Good Views for Contrastive Learning?

- Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, Phillip Isola
- abstract@[open-review](#): Contrastive learning between multiple views of the data has recently achieved state of the art performance in the field of self-supervised representation learning. Despite its success, the influence of different view choices has been less studied. In this paper, we use theoretical and empirical analysis to better understand the importance of view selection, and argue that we should reduce the mutual information (MI) between views while keeping task-relevant information intact. To verify this hypothesis, we devise unsupervised and semi-supervised frameworks that learn effective views by aiming to reduce their MI. We also consider data augmentation as a way to reduce MI, and show that increasing data augmentation indeed leads to decreasing MI and improves downstream classification accuracy. As a by-product, we achieve a new state-of-the-art accuracy on unsupervised pre-training for ImageNet classification (73% top-1 linear readout with a ResNet-50).

Denoising Diffusion Probabilistic Models

- Jonathan Ho, Ajay Jain, Pieter Abbeel
- abstract@[open-review](#): We present high quality image synthesis results using diffusion probabilistic models, a class of latent variable models inspired by considerations from nonequilibrium thermodynamics. Our best results are obtained by training on a weighted variational bound designed according to a novel connection between diffusion probabilistic models and denoising score matching with Langevin dynamics, and our models naturally admit a progressive lossy decompression scheme that can be interpreted as a generalization of autoregressive decoding. On the unconditional CIFAR10 dataset, we obtain an Inception score of 9.46 and a state-of-the-art FID score of 3.17. On 256x256 LSUN, we obtain sample quality similar to ProgressiveGAN.

Barking up the right tree: an approach to search over molecule synthesis DAGs

- John Bradshaw, Brooks Paige, Matt J. Kusner, Marwin Segler, José Miguel Hernández-Lobato
- abstract@[open-review](#): When designing new molecules with particular properties, it is not only important what to make but crucially how to make it. These instructions form a synthesis directed acyclic graph (DAG), describing how a large vocabulary of simple building blocks can be recursively combined through chemical reactions to create more complicated molecules of interest. In contrast, many current deep generative models for molecules ignore synthesizability. We therefore propose a deep generative model that better represents the real world process, by directly outputting molecule synthesis DAGs. We argue that this provides sensible inductive biases, ensuring that our model searches over the same chemical space that chemists would also have access to, as well as interoperability. We show that our approach is able to model chemical space well, producing a wide range of diverse molecules, and allows for unconstrained optimization of an inherently constrained problem: maximize certain chemical properties such that discovered molecules are synthesizable.

[On Uniform Convergence and Low-Norm Interpolation Learning](#)

- Lijia Zhou, Danica J. Sutherland, Nati Srebro
- abstract@[open-review](#): We consider an underdetermined noisy linear regression model where the minimum-norm interpolating predictor is known to be consistent, and ask: can uniform convergence in a norm ball, or at least (following Nagarajan and Kolter) the subset of a norm ball that the algorithm selects on a typical input set, explain this success? We show that uniformly bounding the difference between empirical and population errors cannot show any learning in the norm ball, and cannot show consistency for any set, even one depending on the exact algorithm and distribution. But we argue we can explain the consistency of the minimal-norm interpolator with a slightly weaker, yet standard, notion: uniform convergence of zero-error predictors in a norm ball. We use this to bound the generalization error of low- (but not minimal-)norm interpolating predictors.

[Bandit Samplers for Training Graph Neural Networks](#)

- Ziqi Liu, Zhengwei Wu, Zhiqiang Zhang, Jun Zhou, Shuang Yang, Le Song, Yuan Qi
- abstract@[open-review](#): Several sampling algorithms with variance reduction have been proposed for accelerating the training of Graph Convolution Networks (GCNs). However, due to the intractable computation of optimal sampling distribution, these sampling algorithms are suboptimal for GCNs and are not applicable to more general graph neural networks (GNNs) where the message aggregator contains learned weights rather than fixed weights, such as Graph Attention Networks (GAT). The fundamental reason is that the embeddings of the neighbors or learned weights involved in the optimal sampling distribution are \emph{changing} during the training and \emph{not known a priori}, but only \emph{partially observed} when sampled, thus making the derivation of an optimal variance reduced samplers non-trivial. In this paper, we formulate the optimization of the sampling variance as an adversary bandit problem, where the rewards are related to the node embeddings and learned weights, and can vary constantly. Thus a good sampler needs to acquire variance information about more neighbors (exploration) while at the same time optimizing the immediate sampling variance (exploit). We theoretically show that our algorithm asymptotically approaches the optimal variance within a factor of 3. We show the efficiency and effectiveness of our approach on multiple datasets.

[Sampling from a k-DPP without looking at all items](#)

- Daniele Calandriello, Michal Derezhinski, Michal Valko
- abstract@[open-review](#): Determinantal point processes (DPPs) are a useful probabilistic model for selecting a small diverse subset out of a large collection of items, with applications in summarization, recommendation, stochastic optimization, experimental design and more. Given a kernel function and a subset size k , our goal is to sample k out of n items with probability proportional to the determinant of the kernel matrix induced by the subset (a.k.a. k -DPP). Existing k -DPP sampling algorithms require an expensive preprocessing step which involves multiple passes over all n items, making it infeasible for large datasets. A naïve heuristic addressing this problem is to uniformly subsample a fraction of the data and perform k -DPP sampling only on those items, however this method offers no guarantee that the produced sample will even approximately resemble the target distribution over the original dataset. In this paper, we develop alpha-DPP, an algorithm which adaptively builds a sufficiently large uniform sample of data that is then used to efficiently generate a smaller set of k items, while ensuring that this set is drawn exactly from the target distribution defined on all n items. We show empirically that our algorithm produces a k -DPP sample after observing only a small fraction of all elements, leading to several orders of magnitude faster performance compared to the state-of-the-art. Our implementation of alpha-DPP is provided at <https://github.com/guilgautier/DPPy/>.

[Uncovering the Topology of Time-Varying fMRI Data using Cubical Persistence](#)

- Bastian Rieck, Tristan Yates, Christian Bock, Karsten Borgwardt, Guy Wolf, Nicholas Turk-Browne, Smita Krishnaswamy
- abstract@[open-review](#): Functional magnetic resonance imaging (fMRI) is a crucial technology for gaining insights into cognitive processes in humans. Data amassed from fMRI measurements result in volumetric data sets that vary over time. However, analysing such data presents a challenge due to the large degree of noise and person-to-person variation in how information is represented in the brain. To address this challenge, we present a novel topological approach that encodes each time point in an fMRI data set as a persistence diagram of topological features, i.e. high-dimensional voids present in the data. This representation naturally does not rely on voxel-by-voxel correspondence and is robust towards noise. We show that these time-varying persistence diagrams can be clustered to find meaningful groupings between participants, and that they are also useful in studying within-subject brain state trajectories of subjects performing a particular task. Here, we apply both clustering and trajectory analysis techniques to a group of participants watching the movie 'Partly Cloudy'. We observe significant differences in both brain state trajectories and overall topological activity between adults and children watching the same movie.

[Hierarchical Poset Decoding for Compositional Generalization in Language](#)

- Yinuo Guo, Zeqi Lin, Jian-Guang Lou, Dongmei Zhang
- abstract@[open-review](#): We formalize human language understanding as a structured prediction task where the output is a partially ordered set (poset). Current encoder-decoder architectures do not take the poset structure of semantics into account properly, thus suffering from poor compositional generalization ability. In this paper, we propose a novel hierarchical poset decoding paradigm for compositional generalization in language. Intuitively: (1) the proposed paradigm enforces partial permutation invariance in semantics, thus avoiding overfitting to bias ordering information; (2) the hierarchical mechanism allows to capture high-level structures of posets. We evaluate our proposed decoder on Compositional Freebase Questions (CFQ), a large and realistic natural language question answering dataset that is specifically designed to measure compositional generalization. Results show that it outperforms current decoders.

[Evaluating and Rewarding Teamwork Using Cooperative Game Abstractions](#)

- Tom Yan, Christian Kroer, Alexander Peysakhovich
- abstract@[open-review](#): Can we predict how well a team of individuals will perform together? How should individuals be rewarded for their contributions to the team performance? Cooperative game theory gives us a powerful set of tools for answering these questions: the characteristic function and solution concepts like the Shapley Value. There are two major difficulties in applying these techniques to real world problems: first, the characteristic function is rarely given to us and needs to be learned from data. Second, the Shapley Value is combinatorial in nature. We introduce a parametric model called cooperative game abstractions (CGAs) for estimating characteristic functions from data. CGAs are easy to learn, readily interpretable, and crucially allows linear-time computation of the Shapley Value. We provide identification results and sample complexity bounds for CGA models as well as error bounds in the estimation of the Shapley Value using CGAs. We apply our methods to study teams of artificial RL agents as well as real world teams from professional sports.

[Exchangeable Neural ODE for Set Modeling](#)

- Yang Li, Haidong Yi, Christopher Bender, Siyuan Shan, Junier B. Oliva
- abstract@[open-review](#): Reasoning over an instance composed of a set of vectors, like a point cloud, requires that one accounts for intra-set dependent features among elements. However, since such instances are unordered, the elements' features should remain unchanged when the input's order is permuted. This property, permutation equivariance, is a challenging constraint for most neural architectures. While recent work has proposed global pooling and attention-based solutions, these may be limited in the way that intradependencies are captured in practice. In this work we propose a more

general formulation to achieve permutation equivariance through ordinary differential equations (ODE). Our proposed module, Exchangeable Neural ODE (ExNODE), can be seamlessly applied for both discriminative and generative tasks. We also extend set modeling in the temporal dimension and propose a VAE based model for temporal set modeling. Extensive experiments demonstrate the efficacy of our method over strong baselines.

[Profile Entropy: A Fundamental Measure for the Learnability and Compressibility of Distributions](#)

- Yi Hao, Alon Orlitsky
- abstract@[open-review](#): The profile of a sample is the multiset of its symbol frequencies. We show that for samples of discrete distributions, profile entropy is a fundamental measure unifying the concepts of estimation, inference, and compression. Specifically, profile entropy: a) determines the speed of estimating the distribution relative to the best natural estimator; b) characterizes the rate of inferring all symmetric properties compared with the best estimator over any label-invariant distribution collection; c) serves as the limit of profile compression, for which we derive optimal near-linear-time block and sequential algorithms. To further our understanding of profile entropy, we investigate its attributes, provide algorithms for approximating its value, and determine its magnitude for numerous structural distribution families.

[CoADNet: Collaborative Aggregation-and-Distribution Networks for Co-Salient Object Detection](#)

- Qijian Zhang, Runmin Cong, Junhui Hou, Chongyi Li, Yao Zhao
- abstract@[open-review](#): Co-Salient Object Detection (CoSOD) aims at discovering salient objects that repeatedly appear in a given query group containing two or more relevant images. One challenging issue is how to effectively capture co-saliency cues by modeling and exploiting inter-image relationships. In this paper, we present an end-to-end collaborative aggregation-and-distribution network (CoADNet) to capture both salient and repetitive visual patterns from multiple images. First, we integrate saliency priors into the backbone features to suppress the redundant background information through an online intra-saliency guidance structure. After that, we design a two-stage aggregate-and-distribute architecture to explore group-wise semantic interactions and produce the co-saliency features. In the first stage, we propose a group-attentional semantic aggregation module that models inter-image relationships to generate the group-wise semantic representations. In the second stage, we propose a gated group distribution module that adaptively distributes the learned group semantics to different individuals in a dynamic gating mechanism. Finally, we develop a group consistency preserving decoder tailored for the CoSOD task, which maintains group constraints during feature decoding to predict more consistent full-resolution co-saliency maps. The proposed CoADNet is evaluated on four prevailing CoSOD benchmark datasets, which demonstrates the remarkable performance improvement over ten state-of-the-art competitors.

[Regularized linear autoencoders recover the principal components, eventually](#)

- Xuchan Bao, James Lucas, Sushant Sachdeva, Roger B. Grosse
- abstract@[open-review](#): Our understanding of learning input-output relationships with neural nets has improved rapidly in recent years, but little is known about the convergence of the underlying representations, even in the simple case of linear autoencoders (LAEs). We show that when trained with proper regularization, LAEs can directly learn the optimal representation -- ordered, axis-aligned principal components. We analyze two such regularization schemes: non-uniform L2 regularization and a deterministic variant of nested dropout [Rippel et al, ICML' 2014]. Though both regularization schemes converge to the optimal representation, we show that this convergence is slow due to ill-conditioning that worsens with increasing latent dimension. We show that the inefficiency of learning the optimal representation is not inevitable -- we present a simple modification to the gradient descent update that greatly speeds up convergence empirically.

[Semi-Supervised Partial Label Learning via Confidence-Rated Margin Maximization](#)

- Wei Wang, Min-Ling Zhang
- abstract@[open-review](#): Partial label learning assumes inaccurate supervision where each training example is associated with a set of candidate labels, among which only one is valid. In many real-world scenarios, however, it is costly and time-consuming to assign candidate label sets to all the training examples. To circumvent this difficulty, the problem of semi-supervised partial label learning is investigated in this paper, where unlabeled data is utilized to facilitate model induction along with partial label training examples. Specifically, label propagation is adopted to instantiate the labeling confidence of partial label examples. After that, maximum margin formulation is introduced to jointly enable the induction of predictive model and the estimation of labeling confidence over unlabeled data. The derived formulation enforces confidence-rated margin maximization and confidence manifold preservation over partial label examples and unlabeled data. We show that the predictive model and labeling confidence can be solved via alternating optimization which admits QP solutions in either alternating step. Extensive experiments on synthetic as well as real-world data sets clearly validate the effectiveness of the proposed semi-supervised partial label learning approach.

[GramGAN: Deep 3D Texture Synthesis From 2D Exemplars](#)

- Tiziano Portenier, Siavash Arjomand Bigdeli, Orcun Goksel
- abstract@[open-review](#): We present a novel texture synthesis framework, enabling the generation of infinite, high-quality 3D textures given a 2D exemplar image. Inspired by recent advances in natural texture synthesis, we train deep neural models to generate textures by non-linearly combining learned noise frequencies. To achieve a highly realistic output conditioned on an exemplar patch, we propose a novel loss function that combines ideas from both style transfer and generative adversarial networks. In particular, we train the synthesis network to match the Gram matrices of deep features from a discriminator network. In addition, we propose two architectural concepts and an extrapolation strategy that significantly improve generalization performance. In particular, we inject both model input and condition into hidden network layers by learning to scale and bias hidden activations. Quantitative and qualitative evaluations on a diverse set of exemplars motivate our design decisions and show that our system performs superior to previous state of the art. Finally, we conduct a user study that confirms the benefits of our framework.

[UWSOD: Toward Fully-Supervised-Level Capacity Weakly Supervised Object Detection](#)

- Yunhang Shen, Rongrong Ji, Zhiwei Chen, Yongjian Wu, Feiyue Huang
- abstract@[open-review](#): Weakly supervised object detection (WSOD) has attracted extensive research attention due to its great flexibility of exploiting large-scale dataset with only image-level annotations for detector training. Despite its great advance in recent years, WSOD still suffers limited performance, which is far below that of fully supervised object detection (FSOD). As most WSOD methods depend on object proposal algorithms to generate candidate regions and are also confronted with challenges like low-quality predicted bounding boxes and large scale variation. In this paper, we propose a unified WSOD framework, termed UWSOD, to develop a high-capacity general detection model with only image-level labels, which is self-contained and does not require external modules or additional supervision. To this end, we exploit three important components, i.e., object proposal generation, bounding-box fine-tuning and scale-invariant features. First, we propose an anchor-based self-supervised proposal generator to hypothesize object locations, which is trained end-to-end with supervision created by UWSOD for both objectness classification and regression. Second, we develop a step-wise bounding-box fine-tuning to refine both detection scores and coordinates by progressively select high-confidence object proposals as positive samples, which bootstraps the quality of predicted bounding boxes. Third, we construct a multi-rate resampling pyramid to aggregate multi-scale contextual information, which is the first in-network feature hierarchy to handle scale variation in WSOD. Extensive experiments on PASCAL VOC and MS COCO show that the proposed UWSOD achieves competitive results with the state-of-the-art WSOD methods while not requiring external modules or additional supervision. Moreover, the upper-bound performance of UWSOD with class-agnostic ground-truth bounding boxes approaches Faster R-CNN, which demonstrates UWSOD has fully-supervised-level capacity.

[Learning Restricted Boltzmann Machines with Sparse Latent Variables](#)

- Guy Bresler, Rares-Darius Buhai
- abstract@[open-review](#): Restricted Boltzmann Machines (RBMs) are a common family of undirected graphical models with latent variables. An RBM is described by a bipartite graph, with all observed variables in one layer and all latent variables in the other. We consider the task of learning an RBM given samples generated according to it. The best algorithms for this task currently have time complexity $\tilde{O}(n^2)$ for ferromagnetic RBMs (i.e., with attractive potentials) but $\tilde{O}(n^d)$ for general RBMs, where n is the number of observed variables and d is the maximum degree of a latent variable. Let the MRF neighborhood of an observed variable be its neighborhood in the Markov Random Field of the marginal distribution of the observed variables. In this paper, we give an algorithm for learning general RBMs with time complexity $\tilde{O}(n^{2s+1})$, where s is the maximum number of latent variables connected to the MRF neighborhood of an observed variable. This is an improvement when $s < \log_2(d-1)$, which corresponds to RBMs with sparse latent variables. Furthermore, we give a version of this learning algorithm that recovers a model with small prediction error and whose sample complexity is independent of the minimum potential in the Markov Random Field of the observed variables. This is of interest because the sample complexity of current algorithms scales with the inverse of the minimum potential, which cannot be controlled in terms of natural properties of the RBM.

[Sample Complexity of Asynchronous Q-Learning: Sharper Analysis and Variance Reduction](#)

- Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, Yuxin Chen
- abstract@[open-review](#): Asynchronous Q-learning aims to learn the optimal action-value function (or Q-function) of a Markov decision process (MDP), based on a single trajectory of Markovian samples induced by a behavior policy. Focusing on a γ -discounted MDP with state space S and action space A , we demonstrate that the ℓ_∞ -based sample complexity of classical asynchronous Q-learning --- namely, the number of samples needed to yield an entrywise ϵ -accurate estimate of the Q-function --- is at most on the order of $\frac{1}{\mu_{\min}(1-\gamma)^5} \epsilon^2 + \frac{t_{\text{mix}}}{\mu_{\min}(1-\gamma)} \ln \frac{1}{\epsilon}$ up to some logarithmic factor, provided that a proper constant learning rate is adopted. Here, t_{mix} and μ_{\min} denote respectively the mixing time and the minimum state-action occupancy probability of the sample trajectory. The first term of this bound matches the complexity in the case with independent samples drawn from the stationary distribution of the trajectory. The second term reflects the expense taken for the empirical distribution of the Markovian trajectory to reach a steady state, which is incurred at the very beginning and becomes amortized as the algorithm runs. Encouragingly, the above bound improves upon the state-of-the-art result by a factor of at least $|S||A|$. Further, the scaling on the discount complexity can be improved by means of variance reduction.

[Curriculum learning for multilevel budgeted combinatorial problems](#)

- Adel Nabli, Margarida Carvalho
- abstract@[open-review](#): Learning heuristics for combinatorial optimization problems through graph neural networks have recently shown promising results on some classic NP-hard problems. These are single-level optimization problems with only one player. Multilevel combinatorial optimization problems are their generalization, encompassing situations with multiple players taking decisions sequentially. By framing them in a multi-agent reinforcement learning setting, we devise a value-based method to learn to solve multilevel budgeted combinatorial problems involving two players in a zero-sum game over a graph. Our framework is based on a simple curriculum: if an agent knows how to estimate the value of instances with budgets up to B , then solving instances with budget $B+1$ can be done in polynomial time regardless of the direction of the optimization by checking the value of every possible afterstate. Thus, in a bottom-up approach, we generate datasets of heuristically solved instances with increasingly larger budgets to train our agent. We report results close to optimality on graphs up to 100 nodes and a $185 \times$ speedup on average compared to the quickest exact solver known for the Multilevel Critical Node problem, a max-min-max trilevel problem that has been shown to be at least Σ_2^p -hard.

[FedSplit: an algorithmic framework for fast federated optimization](#)

- Reese Pathak, Martin J. Wainwright
- abstract@[open-review](#): Motivated by federated learning, we consider the hub-and-spoke model of distributed optimization in which a central authority coordinates the computation of a solution among many agents while limiting communication. We first study some past procedures for federated optimization, and show that their fixed points need not correspond to stationary points of the original optimization problem, even in simple convex settings with deterministic updates. In order to remedy these issues, we introduce FedSplit, a class of algorithms based on operator splitting procedures for solving distributed convex minimization with additive structure. We prove that these procedures have the correct fixed points, corresponding to optima of the original optimization problem, and we characterize their convergence rates under different settings. Our theory shows that these methods are provably robust to inexact computation of intermediate local quantities. We complement our theory with some experiments that demonstrate the benefits of our methods in practice.

[Estimation and Imputation in Probabilistic Principal Component Analysis with Missing Not At Random Data](#)

- Aude Sportisse, Claire Boyer, Julie Josse
- abstract@[open-review](#): Missing Not At Random (MNAR) values where the probability of having missing data may depend on the missing value itself, are notoriously difficult to account for in analyses, although very frequent in the data. One solution to handle MNAR data is to specify a model for the missing data mechanism, which makes inference or imputation tasks more complex. Furthermore, this implies a strong a priori on the parametric form of the distribution. However, some works have obtained guarantees on the estimation of parameters in the presence of MNAR data, without specifying the distribution of missing data [\[mohan2018estimation, tang2003analysis\]](#). This is very useful in practice, but is limited to simple cases such as few self-masked MNAR variables in data generated according to linear regression models. We continue this line of research, but extend it to a more general MNAR mechanism, in a more general model of the probabilistic principal component analysis (PPCA), $i.e.$, a low-rank model with random effects. We prove identifiability of the PPCA parameters. We then propose an estimation of the loading coefficients and a data imputation method. They are based on estimators of means, variances and covariances of missing variables, for which consistency is discussed. These estimators have the great advantage of being calculated using only the observed information, leveraging the underlying low-rank structure of the data. We illustrate the relevance of the method with numerical experiments on synthetic data and also on two datasets, one collected from a medical register and the other one from a recommendation system.

[Correlation Robust Influence Maximization](#)

- Louis Chen, Divya Padmanabhan, Chee Chin Lim, Karthik Natarajan
- abstract@[open-review](#): We propose a distributionally robust model for the influence maximization problem. Unlike the classical independent cascade model of Kempe et al (2003), this model's diffusion process is adversarially adapted to the choice of seed set. So instead of optimizing under the assumption that all influence relationships in the network are independent, we seek a seed set whose expected influence under the worst correlation, $i.e.$, the ``worst-case, expected influence'', is maximized. We show that this worst-case influence can be efficiently computed, and though the optimization is NP-hard, a $(1 - 1/e)$ approximation guarantee holds. We also analyze the structure to the adversary's choice of diffusion process, and contrast with established models. Beyond the key computational advantages, we also study the degree to which the independence assumption may be considered costly, and provide insights from numerical experiments comparing the adversarial and independent cascade model.

[Neuronal Gaussian Process Regression](#)

- Johannes Friedrich
- abstract@[open-review](#): The brain takes uncertainty intrinsic to our world into account. For example, associating spatial locations with rewards requires to predict not only expected reward at new spatial locations but also its uncertainty to avoid catastrophic events and forage safely. A powerful and flexible framework for nonlinear regression that takes uncertainty into account in a principled Bayesian manner is Gaussian process (GP) regression. Here I propose that the brain implements GP regression and present neural networks (NNs) for it. First layer neurons, e.g.\ hippocampal place cells, have tuning curves that correspond to evaluations of the GP kernel. Output neurons explicitly and distinctively encode predictive mean and variance, as observed in orbitofrontal cortex (OFC) for the case of reward prediction. Because the weights of a NN implementing exact GP regression do not arise with biological plasticity rules, I present approximations to obtain local (anti-)Hebbian synaptic learning rules. The resulting neuronal network approximates the full GP well compared to popular sparse GP approximations and achieves comparable predictive performance.

[Nonconvex Sparse Graph Learning under Laplacian Constrained Graphical Model](#)

- Jiaxi Ying, JosÃ© VinÃ¡cius de Miranda Cardoso , Daniel Palomar
- abstract@[open-review](#): In this paper, we consider the problem of learning a sparse graph from the Laplacian constrained Gaussian graphical model. This problem can be formulated as a penalized maximum likelihood estimation of the precision matrix under Laplacian structural constraints. Like in the classical graphical lasso problem, recent works made use of the $\|\cdot\|_1$ -norm with the goal of promoting sparsity in the Laplacian constrained precision matrix estimation. However, through empirical evidence, we observe that the $\|\cdot\|_1$ -norm is not effective in imposing a sparse solution in this problem. From a theoretical perspective, we prove that a large regularization parameter will surprisingly lead to a solution representing a fully connected graph instead of a sparse graph. To address this issue, we propose a nonconvex penalized maximum likelihood estimation method, and establish the order of the statistical error. Numerical experiments involving synthetic and real-world data sets demonstrate the effectiveness of the proposed method.

[Synthetic Data Generators -- Sequential and Private](#)

- Olivier Bousquet, Roi Livni, Shay Moran
- abstract@[open-review](#): We study the sample complexity of private synthetic data generation over an unbounded sized class of statistical queries, and show that any class that is privately proper PAC learnable admits a private synthetic data generator (perhaps non-efficient). A differentially private synthetic generator is an algorithm that receives an IID data and publishes synthetic data that is indistinguishable from the true data w.r.t a given fixed class of statistical queries. The synthetic data set can then be used by a data scientist without compromising the privacy of the original data set.

Previous work on synthetic data generators focused on the case that the query class \mathcal{D} is finite and obtained sample complexity bounds that scale logarithmically with the size $|\mathcal{D}|$. Here we construct a private synthetic data generator whose sample complexity is independent of the domain size, and we replace finiteness with the assumption that \mathcal{D} is privately PAC learnable (a formally weaker task, hence we obtain equivalence between the two tasks).

Our proof relies on a new type of synthetic data generator, Sequential Synthetic Data Generators, which we believe may be of interest of their own right. A sequential SDG is defined by a sequential game between a generator that proposes synthetic distributions and a discriminator that tries to distinguish between real and fake distributions. We characterize the classes that admit a sequential-SDG and show that they are exactly Littlestone classes. Given the online nature of the sequential setting, it is natural that Littlestone classes arise in this context. Nevertheless, the characterization of sequential-SDGs by Littlestone classes turns out to be technically challenging, and to the best of the author's knowledge, does not follow via simple reductions to online prediction.

[Uncertainty Quantification for Inferring Hawkes Networks](#)

- Haoyun Wang, Liyan Xie, Alex Cuozzo, Simon Mak, Yao Xie
- abstract@[open-review](#): Multivariate Hawkes processes are commonly used to model streaming networked event data in a wide variety of applications. However, it remains a challenge to extract reliable inference from complex datasets with uncertainty quantification. Aiming towards this, we develop a statistical inference framework to learn causal relationships between nodes from networked data, where the underlying directed graph implies Granger causality. We provide uncertainty quantification for the maximum likelihood estimate of the network multivariate Hawkes process by providing a non-asymptotic confidence set. The main technique is based on the concentration inequalities of continuous-time martingales. We compare our method to the previously-derived asymptotic Hawkes process confidence interval, and demonstrate the strengths of our method in an application to neuronal connectivity reconstruction.

[Implicit Distributional Reinforcement Learning](#)

- Yuguang Yue, Zhendong Wang, Mingyuan Zhou
- abstract@[open-review](#): To improve the sample efficiency of policy-gradient based reinforcement learning algorithms, we propose implicit distributional actor-critic (IDAC) that consists of a distributional critic, built on two deep generator networks (DGNs), and a semi-implicit actor (SIA), powered by a flexible policy distribution. We adopt a distributional perspective on the discounted cumulative return and model it with a state-action-dependent implicit distribution, which is approximated by the DGNs that take state-action pairs and random noises as their input. Moreover, we use the SIA to provide a semi-implicit policy distribution, which mixes the policy parameters with a reparameterizable distribution that is not constrained by an analytic density function. In this way, the policy's marginal distribution is implicit, providing the potential to model complex properties such as covariance structure and skewness, but its parameter and entropy can still be estimated. We incorporate these features with an off-policy algorithm framework to solve problems with continuous action space and compare IDAC with state-of-the-art algorithms on representative OpenAI Gym environments. We observe that IDAC outperforms these baselines in most tasks. Python code is provided.

[Auxiliary Task Reweighting for Minimum-data Learning](#)

- Baifeng Shi, Judy Hoffman, Kate Saenko, Trevor Darrell, Huijuan Xu
- abstract@[open-review](#): Supervised learning requires a large amount of training data, limiting its application where labeled data is scarce. To compensate for data scarcity, one possible method is to utilize auxiliary tasks to provide additional supervision for the main task. Assigning and optimizing the importance weights for different auxiliary tasks remains an crucial and largely understudied research question. In this work, we propose a method to automatically reweight auxiliary tasks in order to reduce the data requirement on the main task. Specifically, we formulate the weighted likelihood function of auxiliary tasks as a surrogate prior for the main task. By adjusting the auxiliary task weights to minimize the divergence between the surrogate prior and the true prior of the main task, we obtain a more accurate prior estimation, achieving the goal of minimizing the required amount of training data for the main task and avoiding a costly grid search. In multiple experimental settings (e.g. semi-supervised learning, multi-label classification), we demonstrate that our algorithm can effectively utilize limited labeled data of the main task with the benefit of auxiliary tasks compared with previous task reweighting methods. We also show that under extreme cases with only a few extra examples (e.g. few-shot domain adaptation), our algorithm results in significant improvement over the baseline. Our code and video is available at <https://sites.google.com/view/auxiliary-task-reweighting>.

[Small Nash Equilibrium Certificates in Very Large Games](#)

- Brian Zhang, Tuomas Sandholm
- abstract@[open-review](#): In many game settings, the game is not explicitly given but is only accessible by playing it. While there have been impressive demonstrations in such settings, prior techniques have not offered safety guarantees, that is, guarantees on the game-theoretic exploitability of the

computed strategies. In this paper we introduce an approach that shows that it is possible to provide exploitability guarantees in such settings without ever exploring the entire game. We introduce a notion of a certificate of an extensive-form approximate Nash equilibrium. For verifying a certificate, we give an algorithm that runs in time linear in the size of the certificate rather than the size of the whole game. In zero-sum games, we further show that an optimal certificate---given the exploration so far---can be computed with any standard game-solving algorithm (e.g., using a linear program or counterfactual regret minimization). However, unlike in the cases of normal form or perfect information, we show that certain families of extensive-form games do not have small approximate certificates, even after making extremely nice assumptions on the structure of the game. Despite this difficulty, we find experimentally that very small certificates, even exact ones, often exist in large and even in infinite games. Overall, our approach enables one to try one's favorite exploration strategies while offering exploitability guarantees, thereby decoupling the exploration strategy from the equilibrium-finding process.

[Training Linear Finite-State Machines](#)

- Arash Ardakani, Amir Ardakani, Warren Gross
- abstract@[open-review](#): A finite-state machine (FSM) is a computation model to process binary strings in sequential circuits. Hence, a single-input linear FSM is conventionally used to implement complex single-input functions , such as tanh and exponentiation functions, in stochastic computing (SC) domain where continuous values are represented by sequences of random bits. In this paper, we introduce a method that can train a multi-layer FSM-based network where FSMs are connected to every FSM in the previous and the next layer. We show that the proposed FSM-based network can synthesize multi-input complex functions such as 2D Gabor filters and can perform non-sequential tasks such as image classifications on stochastic streams with no multiplication since FSMs are implemented by look-up tables only. Inspired by the capability of FSMs in processing binary streams, we then propose an FSM-based model that can process time series data when performing temporal tasks such as character-level language modeling. Unlike long short-term memories (LSTMs) that unroll the network for each input time step and perform back-propagation on the unrolled network, our FSM-based model requires to backpropagate gradients only for the current input time step while it is still capable of learning long-term dependencies. Therefore, our FSM-based model can learn extremely long-term dependencies as it requires $1/l$ memory storage during training compared to LSTMs, where l is the number of time steps. Moreover, our FSM-based model reduces the power consumption of training on a GPU by 33% compared to an LSTM model of the same size.

[Efficient active learning of sparse halfspaces with arbitrary bounded noise](#)

- Chicheng Zhang, Jie Shen, Pranjal Awasthi
- abstract@[open-review](#): We study active learning of homogeneous $\$s\$$ -sparse halfspaces in $\$mathbb{R}^d\$$ under the setting where the unlabeled data distribution is isotropic log-concave and each label is flipped with probability at most $\$\\eta\$$ for a parameter $\$\\eta \\in [0, \\frac{1}{2}]\$$, known as the bounded noise. Even in the presence of mild label noise, i.e. $\$\\eta\$$ is a small constant, this is a challenging problem and only recently have label complexity bounds of the form $\$\\tilde{O}(s \\cdot polylog(d, \\frac{1}{\\epsilon}))\$$ been established in [Zhang 2018] for computationally efficient algorithms. In contrast, under high levels of label noise, the label complexity bounds achieved by computationally efficient algorithms are much worse: the best known result [Awasthi et al. 2016] provides a computationally efficient algorithm with label complexity $\$\\tilde{O}((s \\ln d/\\epsilon)^{poly(1/(1-2\\eta))})\$$, which is label-efficient only when the noise rate $\$\\eta\$$ is a fixed constant. In this work, we substantially improve on it by designing a polynomial time algorithm for active learning of $\$s\$$ -sparse halfspaces, with a label complexity of $\$\\tilde{O}(\\frac{s}{(1-2\\eta)^4} \\cdot polylog(d, \\frac{1}{\\epsilon}))\$$. This is the first efficient algorithm with label complexity polynomial in $\$\\frac{1}{(1-2\\eta)}\$$ in this setting, which is label-efficient even for $\$\\eta\$$ arbitrarily close to $\$\\frac{1}{2}\$$. Our active learning algorithm and its theoretical guarantees also immediately translate to new state-of-the-art label and sample complexity results for full-dimensional active and passive halfspace learning under arbitrary bounded noise.

[Swapping Autoencoder for Deep Image Manipulation](#)

- Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, Richard Zhang
- abstract@[open-review](#): Deep generative models have become increasingly effective at producing realistic images from randomly sampled seeds, but using such models for controllable manipulation of existing images remains challenging. We propose the Swapping Autoencoder, a deep model designed specifically for image manipulation, rather than random sampling. The key idea is to encode an image into two independent components and enforce that any swapped combination maps to a realistic image. In particular, we encourage the components to represent structure and texture, by enforcing one component to encode co-occurring patch statistics across different parts of the image. As our method is trained with an encoder, finding the latent codes for a new input image becomes trivial, rather than cumbersome. As a result, our method enables us to manipulate real input images in various ways, including texture swapping, local and global editing, and latent code vector arithmetic. Experiments on multiple datasets show that our model produces better results and is substantially more efficient compared to recent generative models.

[Self-Supervised Few-Shot Learning on Point Clouds](#)

- Charu Sharma, Manohar Kaul
- abstract@[open-review](#): The increased availability of massive point clouds coupled with their utility in a wide variety of applications such as robotics, shape synthesis, and self-driving cars has attracted increased attention from both industry and academia. Recently, deep neural networks operating on labeled point clouds have shown promising results on supervised learning tasks like classification and segmentation. However, supervised learning leads to the cumbersome task of annotating the point clouds. To combat this problem, we propose two novel self-supervised pre-training tasks that encode a hierarchical partitioning of the point clouds using a cover-tree, where point cloud subsets lie within balls of varying radii at each level of the cover-tree. Furthermore, our self-supervised learning network is restricted to pre-train on the support set (comprising of scarce training examples) used to train the downstream network in a few-shot learning (FSL) setting. Finally, the fully-trained self-supervised network's point embeddings are input to the downstream task's network. We present a comprehensive empirical evaluation of our method on both downstream classification and segmentation tasks and show that supervised methods pre-trained with our self-supervised learning method significantly improve the accuracy of state-of-the-art methods. Additionally, our method also outperforms previous unsupervised methods in downstream classification tasks.

[Faster Differentially Private Samplers via RĂ©nyi Divergence Analysis of Discretized Langevin MCMC](#)

- Arun Ganesh, Kunal Talwar
- abstract@[open-review](#): Various differentially private algorithms instantiate the exponential mechanism, and require sampling from the distribution $\$exp(-f)\$$ for a suitable function $\$f\$$. When the domain of the distribution is high-dimensional, this sampling can be challenging. Using heuristic sampling schemes such as Gibbs sampling does not necessarily lead to provable privacy. When $\$f\$$ is convex, techniques from log-concave sampling lead to polynomial-time algorithms, albeit with large polynomials. Langevin dynamics-based algorithms offer much faster alternatives under some distance measures such as statistical distance. In this work, we establish rapid convergence for these algorithms under distance measures more suitable for differential privacy. For smooth, strongly-convex $\$f\$$, we give the first results proving convergence in RĂ©nyi divergence. This gives us fast differentially private algorithms for such $\$f\$$. Our techniques are simple and generic and apply also to underdamped Langevin dynamics.

[Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-VAE](#)

- Ding Zhou, Xue-Xin Wei
- abstract@[open-review](#): The ability to record activities from hundreds of neurons simultaneously in the brain has placed an increasing demand for developing appropriate statistical techniques to analyze such data. Recently, deep generative models have been proposed to fit neural population

responses. While these methods are flexible and expressive, the downside is that they can be difficult to interpret and identify. To address this problem, we propose a method that integrates key ingredients from latent models and traditional neural encoding models. Our method, pi-VAE, is inspired by recent progress on identifiable variational auto-encoder, which we adapt to make appropriate for neuroscience applications. Specifically, we propose to construct latent variable models of neural activity while simultaneously modeling the relation between the latent and task variables (non-neural variables, e.g. sensory, motor, and other externally observable states). The incorporation of task variables results in models that are not only more constrained, but also show qualitative improvements in interpretability and identifiability. We validate pi-VAE using synthetic data, and apply it to analyze neurophysiological datasets from rat hippocampus and macaque motor cortex. We demonstrate that pi-VAE not only fits the data better, but also provides unexpected novel insights into the structure of the neural codes.

[RL Unplugged: A Suite of Benchmarks for Offline Reinforcement Learning](#)

- Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Thomas Paine, Sergio GÃ³mez, Konrad Zolna, Rishabh Agarwal, Josh S. Merel, Daniel J. Mankowitz, Cosmin Paduraru, Gabriel Dulac-Arnold, Jerry Li, Mohammad Norouzi, Matthew Hoffman, Nicolas Heess, Nando de Freitas
- abstract@[open-review](#): Offline methods for reinforcement learning have a potential to help bridge the gap between reinforcement learning research and real-world applications. They make it possible to learn policies from offline datasets, thus overcoming concerns associated with online data collection in the real-world, including cost, safety, or ethical concerns. In this paper, we propose a benchmark called RL Unplugged to evaluate and compare offline RL methods. RL Unplugged includes data from a diverse range of domains including games e.g., Atari benchmark) and simulated motor control problems (e.g., DM Control Suite). The datasets include domains that are partially or fully observable, use continuous or discrete actions, and have stochastic vs. deterministic dynamics. We propose detailed evaluation protocols for each domain in RL Unplugged and provide an extensive analysis of supervised learning and offline RL methods using these protocols. We will release data for all our tasks and open-source all algorithms presented in this paper. We hope that our suite of benchmarks will increase the reproducibility of experiments and make it possible to study challenging tasks with a limited computational budget, thus making RL research both more systematic and more accessible across the community. Moving forward, we view RL Unplugged as a living benchmark suite that will evolve and grow with datasets contributed by the research community and ourselves. Our project page is available on github.

[Dual T: Reducing Estimation Error for Transition Matrix in Label-noise Learning](#)

- Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, Masashi Sugiyama
- abstract@[open-review](#): The transition matrix, denoting the transition relationship from clean labels to noisy labels, is essential to build statistically consistent classifiers in label-noise learning. Existing methods for estimating the transition matrix rely heavily on estimating the noisy class posterior. However, the estimation error for noisy class posterior could be large because of the randomness of label noise. The estimation error would lead the transition matrix to be poorly estimated. Therefore in this paper, we aim to solve this problem by exploiting the divide-and-conquer paradigm. Specifically, we introduce an intermediate class to avoid directly estimating the noisy class posterior. By this intermediate class, the original transition matrix can then be factorized into the product of two easy-to-estimated transition matrices. We term the proposed method as the dual \$T\$-estimator. Both theoretical analyses and empirical results illustrate the effectiveness of the dual \$T\$-estimator for estimating transition matrices, leading to better classification performances.

[Interior Point Solving for LP-based prediction+optimisation](#)

- Jayanta Mandi, Tias Guns
- abstract@[open-review](#): Solving optimization problem is the key to decision making in many real-life analytics applications. However, the coefficients of the optimization problems are often uncertain and dependent on external factors, such as future demand or energy- or stock prices. Machine learning (ML) models, especially neural networks, are increasingly being used to estimate these coefficients in a data-driven way. Hence, end-to-end predict-and-optimize approaches, which consider how effective the predicted values are to solve the optimization problem, have received increasing attention. In case of integer linear programming problems, a popular approach to overcome their non-differentiability is to add a quadratic penalty term to the continuous relaxation, such that results from differentiating over quadratic programs can be used. Instead we investigate the use of the more principled logarithmic barrier term, as widely used in interior point solvers for linear programming. Instead of differentiating the KKT conditions, we consider the homogeneous self-dual formulation of the LP and we show the relation between the interior point step direction and corresponding gradients needed for learning. Finally, our empirical experiments demonstrate our approach performs as good as if not better than the state-of-the-art QPTL (Quadratic Programming task loss) formulation of Wilder et al. and SPO approach of Elmachtoub and Grigas.

[A simple normative network approximates local non-Hebbian learning in the cortex](#)

- Siavash Golkar, David Lipshutz, Yanis Bahroun, Anirvan Sengupta, Dmitri Chklovskii
- abstract@[open-review](#): To guide behavior, the brain extracts relevant features from high-dimensional data streamed by sensory organs. Neuroscience experiments demonstrate that the processing of sensory inputs by cortical neurons is modulated by instructive signals which provide context and task-relevant information. Here, adopting a normative approach, we model these instructive signals as supervisory inputs guiding the projection of the feedforward data. Mathematically, we start with a family of Reduced-Rank Regression (RRR) objective functions which include Reduced Rank (minimum) Mean Square Error (RRMSE) and Canonical Correlation Analysis (CCA), and derive novel offline and online optimization algorithms, which we call Bio-RRR. The online algorithms can be implemented by neural networks whose synaptic learning rules resemble calcium plateau potential dependent plasticity observed in the cortex. We detail how, in our model, the calcium plateau potential can be interpreted as a backpropagating error signal. We demonstrate that, despite relying exclusively on biologically plausible local learning rules, our algorithms perform competitively with existing implementations of RRMSE and CCA.

[Kernelized information bottleneck leads to biologically plausible 3-factor Hebbian learning in deep networks](#)

- Roman Pogodin, Peter Latham
- abstract@[open-review](#): The state-of-the art machine learning approach to training deep neural networks, backpropagation, is implausible for real neural networks: neurons need to know their outgoing weights; training alternates between a bottom-up forward pass (computation) and a top-down backward pass (learning); and the algorithm often needs precise labels of many data points. Biologically plausible approximations to backpropagation, such as feedback alignment, solve the weight transport problem, but not the other two. Thus, fully biologically plausible learning rules have so far remained elusive. Here we present a family of learning rules that does not suffer from any of these problems. It is motivated by the information bottleneck principle (extended with kernel methods), in which networks learn to compress the input as much as possible without sacrificing prediction of the output. The resulting rules have a 3-factor Hebbian structure: they require pre- and post-synaptic firing rates and an error signal - the third factor - consisting of a global teaching signal and a layer-specific term, both available without a top-down pass. They do not require precise labels; instead, they rely on the similarity between pairs of desired outputs. Moreover, to obtain good performance on hard problems and retain biological plausibility, our rules need divisive normalization - a known feature of biological networks. Finally, simulations show that our rules perform nearly as well as backpropagation on image classification tasks.

[Understanding the Role of Training Regimes in Continual Learning](#)

- Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, Hassan Ghasemzadeh

- abstract@[open-review](#): Catastrophic forgetting affects the training of neural networks, limiting their ability to learn multiple tasks sequentially. From the perspective of the well established plasticity-stability dilemma, neural networks tend to be overly plastic, lacking the stability necessary to prevent the forgetting of previous knowledge, which means that as learning progresses, networks tend to forget previously seen tasks. This phenomenon coined in the continual learning literature, has attracted much attention lately, and several families of approaches have been proposed with different degrees of success. However, there has been limited prior work extensively analyzing the impact that different training regimes -- learning rate, batch size, regularization method-- can have on forgetting. In this work, we depart from the typical approach of altering the learning algorithm to improve stability. Instead, we hypothesize that the geometrical properties of the local minima found for each task play an important role in the overall degree of forgetting. In particular, we study the effect of dropout, learning rate decay, and batch size on forming training regimes that widen the tasks' local minima and consequently, on helping it not to forget catastrophically. Our study provides practical insights to improve stability via simple yet effective techniques that outperform alternative baselines.

[Fair regression with Wasserstein barycenters](#)

- Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, Massimiliano Pontil
- abstract@[open-review](#): We study the problem of learning a real-valued function that satisfies the Demographic Parity constraint. It demands the distribution of the predicted output to be independent of the sensitive attribute. We consider the case that the sensitive attribute is available for prediction. We establish a connection between fair regression and optimal transport theory, based on which we derive a close form expression for the optimal fair predictor. Specifically, we show that the distribution of this optimum is the Wasserstein barycenter of the distributions induced by the standard regression function on the sensitive groups. This result offers an intuitive interpretation of the optimal fair prediction and suggests a simple post-processing algorithm to achieve fairness. We establish risk and distribution-free fairness guarantees for this procedure. Numerical experiments indicate that our method is very effective in learning fair models, with a relative increase in error rate that is inferior to the relative gain in fairness.

[Training Stronger Baselines for Learning to Optimize](#)

- Tianlong Chen, Weiyi Zhang, Zhou Jingyang, Shiyu Chang, Sijia Liu, Lisa Amini, Zhangyang Wang
- abstract@[open-review](#): Learning to optimize (L2O) is gaining increased attention because classical optimizers require laborious, problem-specific design and hyperparameter tuning. However, there are significant performance and practicality gaps between manually designed optimizers and existing L2O models. Specifically, learned optimizers are applicable to only a limited class of problems, often exhibit instability, and generalize poorly. As research efforts focus on increasingly sophisticated L2O models, we argue for an orthogonal, under-explored theme: improved training techniques for L2O models. We first present a progressive, curriculum-based training scheme, which gradually increases the optimizer unroll length to mitigate the well-known L2O dilemma of truncation bias (shorter unrolling) versus gradient explosion (longer unrolling). Secondly, we present an off-policy imitation learning based approach to guide the L2O learning, by learning from the behavior of analytical optimizers. We evaluate our improved training techniques with a variety of state-of-the-art L2O models and immediately boost their performance, without making any change to their model structures. We demonstrate that, using our improved training techniques, one of the earliest and simplest L2O models can be trained to outperform even the latest and most complex L2O models on a number of tasks. Our results demonstrate a greater potential of L2O yet to be unleashed, and prompt a reconsideration of recent L2O model progress. Our codes are publicly available at: <https://github.com/VITA-Group/L2O-Training-Techniques>.

[Exactly Computing the Local Lipschitz Constant of ReLU Networks](#)

- Matt Jordan, Alexandros G. Dimakis
- abstract@[open-review](#): The local Lipschitz constant of a neural network is a useful metric with applications in robustness, generalization, and fairness evaluation. We provide novel analytic results relating the local Lipschitz constant of nonsmooth vector-valued functions to a maximization over the norm of the generalized Jacobian. We present a sufficient condition for which backpropagation always returns an element of the generalized Jacobian, and reframe the problem over this broad class of functions. We show strong inapproximability results for estimating Lipschitz constants of ReLU networks, and then formulate an algorithm to compute these quantities exactly. We leverage this algorithm to evaluate the tightness of competing Lipschitz estimators and the effects of regularized training on the Lipschitz constant.

[Strictly Batch Imitation Learning by Energy-based Distribution Matching](#)

- Daniel Jarrett, Ioana Bica, Mihaela van der Schaar
- abstract@[open-review](#): Consider learning a policy purely on the basis of demonstrated behavior---that is, with no access to reinforcement signals, no knowledge of transition dynamics, and no further interaction with the environment. This strictly batch imitation learning problem arises wherever live experimentation is costly, such as in healthcare. One solution is simply to retrofit existing algorithms for apprenticeship learning to work in the offline setting. But such an approach leans heavily on off-policy evaluation or offline model estimation, and can be indirect and inefficient. We argue that a good solution should be able to explicitly parameterize a policy (i.e. respecting action conditionals), implicitly learn from rollout dynamics (i.e. leveraging state marginals), and---crucially---operate in an entirely offline fashion. To address this challenge, we propose a novel technique by energy-based distribution matching (EDM): By identifying parameterizations of the (discriminative) model of a policy with the (generative) energy function for state distributions, EDM yields a simple but effective solution that equivalently minimizes a divergence between the occupancy measure for the demonstrator and a model thereof for the imitator. Through experiments with application to control and healthcare settings, we illustrate consistent performance gains over existing algorithms for strictly batch imitation learning.

[On the Ergodicity, Bias and Asymptotic Normality of Randomized Midpoint Sampling Method](#)

- Ye He, Krishnakumar Balasubramanian, Murat A. Erdogdu
- abstract@[open-review](#): The randomized midpoint method, proposed by (Shen and Lee, 2019), has emerged as an optimal discretization procedure for simulating the continuous time underdamped Langevin diffusion. In this paper, we analyze several probabilistic properties of the randomized midpoint discretization method, considering both overdamped and underdamped Langevin dynamics. We first characterize the stationary distribution of the discrete chain obtained with constant step-size discretization and show that it is biased away from the target distribution. Notably, the step-size needs to go to zero to obtain asymptotic unbiasedness. Next, we establish the asymptotic normality of numerical integration using the randomized midpoint method and highlight the relative advantages and disadvantages over other discretizations. Our results collectively provide several insights into the behavior of the randomized midpoint discretization method, including obtaining confidence intervals for numerical integrations.

[A Single-Loop Smoothed Gradient Descent-Ascent Algorithm for Nonconvex-Concave Min-Max Problems](#)

- Jiawei Zhang, Peijun Xiao, Ruoyu Sun, Zhiqian Luo
- abstract@[open-review](#): Nonconvex-concave min-max problem arises in many machine learning applications including minimizing a pointwise maximum of a set of nonconvex functions and robust adversarial training of neural networks. A popular approach to solve this problem is the gradient descent-ascent (GDA) algorithm which unfortunately can exhibit oscillation in case of nonconvexity. In this paper, we introduce a ``smoothing'' scheme which can be combined with GDA to stabilize the oscillation and ensure convergence to a stationary solution. We prove that the stabilized GDA algorithm can achieve an $\mathcal{O}(1/\epsilon^2)$ iteration complexity for minimizing the pointwise maximum of a finite collection of nonconvex functions. Moreover, the smoothed GDA algorithm achieves an $\mathcal{O}(1/\epsilon^4)$ iteration complexity for general nonconvex-concave problems. Extensions of this stabilized GDA

algorithm to multi-block cases are presented. To the best of our knowledge, this is the first algorithm to achieve $\mathcal{O}(1/\epsilon^2)$ for a class of nonconvex-concave problem. We illustrate the practical efficiency of the stabilized GDA algorithm on robust training.

[Generating Correct Answers for Progressive Matrices Intelligence Tests](#)

- Niv Pekar, Yaniv Benny, Lior Wolf
- abstract@[open-review](#): Ravenâ€™s Progressive Matrices are multiple-choice intelligence tests, where one tries to complete the missing location in a 3x3 grid of abstract images. Previous attempts to address this test have focused solely on selecting the right answer out of the multiple choices. In this work, we focus, instead, on generating a correct answer given the grid, which is a harder task, by definition. The proposed neural model combines multiple advances in generative models, including employing multiple pathways through the same network, using the reparameterization trick along two pathways to make their encoding compatible, a selective application of variational losses, and a complex perceptual loss that is coupled with a selective backpropagation procedure. Our algorithm is able not only to generate a set of plausible answers but also to be competitive to the state of the art methods in multiple-choice tests.

[HyNet: Learning Local Descriptor with Hybrid Similarity Measure and Triplet Loss](#)

- Yurun Tian, Axel Barroso Laguna, Tony Ng, Vassileios Balntas, Krystian Mikolajczyk
- abstract@[open-review](#): In this paper, we investigate how L2 normalisation affects the back-propagated descriptor gradients during training. Based on our observations, we propose HyNet, a new local descriptor that leads to state-of-the-art results in matching. HyNet introduces a hybrid similarity measure for triplet margin loss, a regularisation term constraining the descriptor norm, and a new network architecture that performs L2 normalisation of all intermediate feature maps and the output descriptors. HyNet surpasses previous methods by a significant margin on standard benchmarks that include patch matching, verification, and retrieval, as well as outperforming full end-to-end methods on 3D reconstruction tasks.

[Preference learning along multiple criteria: A game-theoretic perspective](#)

- Kush Bhatia, Ashwin Pananjady, Peter Bartlett, Anca Dragan, Martin J. Wainwright
- abstract@[open-review](#): The literature on ranking from ordinal data is vast, and there are several ways to aggregate overall preferences from pairwise comparisons between objects. In particular, it is well-known that any Nash equilibrium of the zero-sum game induced by the preference matrix defines a natural solution concept (winning distribution over objects) known as a von Neumann winner. Many real-world problems, however, are inevitably multi-criteria, with different pairwise preferences governing the different criteria. In this work, we generalize the notion of a von Neumann winner to the multi-criteria setting by taking inspiration from Blackwellâ€™s approachability. Our framework allows for non-linear aggregation of preferences across criteria, and generalizes the linearization-based approach from multi-objective optimization. From a theoretical standpoint, we show that the Blackwell winner of a multi-criteria problem instance can be computed as the solution to a convex optimization problem. Furthermore, given random samples of pairwise comparisons, we show that a simple, "plug-in" estimator achieves (near-)optimal minimax sample complexity. Finally, we showcase the practical utility of our framework in a user study on autonomous driving, where we find that the Blackwell winner outperforms the von Neumann winner for the overall preferences.

[Multi-Plane Program Induction with 3D Box Priors](#)

- Yikai Li, Jiayuan Mao, Xiuming Zhang, Bill Freeman, Josh Tenenbaum, Noah Snavely, Jiajun Wu
- abstract@[open-review](#): We consider two important aspects in understanding and editing images: modeling regular, program-like texture or patterns in 2D planes, and 3D posing of these planes in the scene. Unlike prior work on image-based program synthesis, which assumes the image contains a single visible 2D plane, we present Box Program Induction (BPI), which infers a program-like scene representation that simultaneously models repeated structure on multiple 2D planes, the 3D position and orientation of the planes, and camera parameters, all from a single image. Our model assumes a box prior, i.e., that the image captures either an inner view or an outer view of a box in 3D. It uses neural networks to infer visual cues such as vanishing points, wireframe lines to guide a search-based algorithm to find the program that best explains the image. Such a holistic, structured scene representation enables 3D-aware interactive image editing operations such as inpainting missing pixels, changing camera parameters, and extrapolate the image contents.

[Online Neural Connectivity Estimation with Noisy Group Testing](#)

- Anne Draelos, John Pearson
- abstract@[open-review](#): One of the primary goals of systems neuroscience is to relate the structure of neural circuits to their function, yet patterns of connectivity are difficult to establish when recording from large populations in behaving organisms. Many previous approaches have attempted to estimate functional connectivity between neurons using statistical modeling of observational data, but these approaches rely heavily on parametric assumptions and are purely correlational. Recently, however, holographic photostimulation techniques have made it possible to precisely target selected ensembles of neurons, offering the possibility of establishing direct causal links. A naive method for inferring functional connections is to stimulate each individual neuron multiple times and observe the responses of cells in the local network, but this approach scales poorly with the number of neurons. Here, we propose a method based on noisy group testing that drastically increases the efficiency of this process in sparse networks. By stimulating small ensembles of neurons, we show that it is possible to recover binarized network connectivity with a number of tests that grows only logarithmically with population size under minimal statistical assumptions. Moreover, we prove that our approach, which reduces to an efficiently solvable convex optimization problem, can be related to Variational Bayesian inference on the binary connection weights, and we derive rigorous bounds on the posterior marginals. This allows us to extend our method to the streaming setting, where continuously updated posteriors allow for optional stopping, and we demonstrate the feasibility of inferring connectivity for networks of up to tens of thousands of neurons online.

[Once-for-All Adversarial Training: In-Situ Tradeoff between Robustness and Accuracy for Free](#)

- Haotao Wang, Tianlong Chen, Shupeng Gui, TingKuei Hu, Ji Liu, Zhangyang Wang
- abstract@[open-review](#): Adversarial training and its many variants substantially improve deep network robustness, yet at the cost of compromising standard accuracy. Moreover, the training process is heavy and hence it becomes impractical to thoroughly explore the trade-off between accuracy and robustness. This paper asks this new question: how to quickly calibrate a trained model in-situ, to examine the achievable trade-offs between its standard and robust accuracies, without (re-)training it many times? Our proposed framework, Once-for-all Adversarial Training (OAT), is built on an innovative model-conditional training framework, with a controlling hyper-parameter as the input. The trained model could be adjusted among different standard and robust accuracies â€œfor freeâ€ at testing time. As an important knob, we exploit dual batch normalization to separate standard and adversarial feature statistics, so that they can be learned in one model without degrading performance. We further extend OAT to a Once-for-all Adversarial Training and Slimming (OATS) framework, that allows for the joint trade-off among accuracy, robustness and runtime efficiency. Experiments show that, without any re-training nor ensembling, OAT/OATS achieve similar or even superior performance compared to dedicatedly trained models at various configurations. Our codes and pretrained models are available at: <https://github.com/VITA-Group/Once-for-All-Adversarial-Training>.

[Implicit Neural Representations with Periodic Activation Functions](#)

- Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, Gordon Wetzstein

- abstract@[open-review](#): Implicitly defined, continuous, differentiable signal representations parameterized by neural networks have emerged as a powerful paradigm, offering many possible benefits over conventional representations. However, current network architectures for such implicit neural representations are incapable of modeling signals with fine detail, and fail to represent a signal's spatial and temporal derivatives, despite the fact that these are essential to many physical signals defined implicitly as the solution to partial differential equations. We propose to leverage periodic activation functions for implicit neural representations and demonstrate that these networks, dubbed sinusoidal representation networks or SIRENs, are ideally suited for representing complex natural signals and their derivatives. We analyze SIREN activation statistics to propose a principled initialization scheme and demonstrate the representation of images, wavefields, video, sound, and their derivatives. Further, we show how SIRENs can be leveraged to solve challenging boundary value problems, such as particular Eikonal equations (yielding signed distance functions), the Poisson equation, and the Helmholtz and wave equations. Lastly, we combine SIRENs with hypernetworks to learn priors over the space of SIREN functions.

[Rotated Binary Neural Network](#)

- Mingbao Lin, Rongrong Ji, Zihan Xu, Baochang Zhang, Yan Wang, Yongjian Wu, Feiyue Huang, Chia-Wen Lin
- abstract@[open-review](#): Binary Neural Network (BNN) shows its predominance in reducing the complexity of deep neural networks. However, it suffers severe performance degradation. One of the major impediments is the large quantization error between the full-precision weight vector and its binary vector. Previous works focus on compensating for the norm gap while leaving the angular bias hardly touched. In this paper, for the first time, we explore the influence of angular bias on the quantization error and then introduce a Rotated Binary Neural Network (RBNN), which considers the angle alignment between the full-precision weight vector and its binarized version. At the beginning of each training epoch, we propose to rotate the full-precision weight vector to its binary vector to reduce the angular bias. To avoid the high complexity of learning a large rotation matrix, we further introduce a bi-rotation formulation that learns two smaller rotation matrices. In the training stage, we devise an adjustable rotated weight vector for binarization to escape the potential local optimum. Our rotation leads to around 50% weight flips which maximize the information gain. Finally, we propose a training-aware approximation of the sign function for the gradient backward. Experiments on CIFAR-10 and ImageNet demonstrate the superiorities of RBNN over many state-of-the-arts. Our source code, experimental settings, training logs and binary models are available at <https://github.com/lmbxmu/RBNN>.

[Community detection in sparse time-evolving graphs with a dynamical Bethe-Hessian](#)

- Lorenzo Dall'Amico, Romain Couillet, Nicolas Tremblay
- abstract@[open-review](#): This article considers the problem of community detection in sparse dynamical graphs in which the community structure evolves over time. A fast spectral algorithm based on an extension of the Bethe-Hessian matrix is proposed, which benefits from the positive correlation in the class labels and in their temporal evolution and is designed to be applicable to any dynamical graph with a community structure. Under the dynamical degree-corrected stochastic block model, in the case of two classes of equal size, we demonstrate and support with extensive simulations that our proposed algorithm is capable of making non-trivial community reconstruction as soon as theoretically possible, thereby reaching the optimal detectability threshold and provably outperforming competing spectral methods.

[Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness](#)

- Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, Balaji Lakshminarayanan
- abstract@[open-review](#): Bayesian neural networks (BNN) and deep ensembles are principled approaches to estimate the predictive uncertainty of a deep learning model. However their practicality in real-time, industrial-scale applications are limited due to their heavy memory and inference cost. This motivates us to study principled approaches to high-quality uncertainty estimation that require only a single deep neural network (DNN). By formalizing the uncertainty quantification as a minimax learning problem, we first identify input distance awareness, i.e., the model's ability to quantify the distance of a testing example from the training data in the input space, as a necessary condition for a DNN to achieve high-quality (i.e., minimax optimal) uncertainty estimation. We then propose Spectral-normalized Neural Gaussian Process (SNGP), a simple method that improves the distance-awareness ability of modern DNNs, by adding a weight normalization step during training and replacing the output layer. On a suite of vision and language understanding tasks and on modern architectures (Wide-ResNet and BERT), SNGP is competitive with deep ensembles in prediction, calibration and out-of-domain detection, and outperforms the other single-model approaches.

[Adaptive Learning of Rank-One Models for Efficient Pairwise Sequence Alignment](#)

- Govinda Kamath, Tavor Baharav, Ilan Shomorony
- abstract@[open-review](#): Pairwise alignment of DNA sequencing data is a ubiquitous task in bioinformatics and typically represents a heavy computational burden. State-of-the-art approaches to speed up this task use hashing to identify short segments (k-mers) that are shared by pairs of reads, which can then be used to estimate alignment scores. However, when the number of reads is large, accurately estimating alignment scores for all pairs is still very costly. Moreover, in practice, one is only interested in identifying pairs of reads with large alignment scores. In this work, we propose a new approach to pairwise alignment estimation based on two key new ingredients. The first ingredient is to cast the problem of pairwise alignment estimation under a general framework of rank-one crowdsourcing models, where the workers' responses correspond to k-mer hash collisions. These models can be accurately solved via a spectral decomposition of the response matrix. The second ingredient is to utilise a multi-armed bandit algorithm to adaptively refine this spectral estimator only for read pairs that are likely to have large alignments. The resulting algorithm iteratively performs a spectral decomposition of the response matrix for adaptively chosen subsets of the read pairs.

[Hierarchical nucleation in deep neural networks](#)

- Diego Doimo, Aldo Glielmo, Alessio Ansuini, Alessandro Laio
- abstract@[open-review](#): Deep convolutional networks (DCNs) learn meaningful representations where data that share the same abstract characteristics are positioned closer and closer. Understanding these representations and how they are generated is of unquestioned practical and theoretical interest. In this work we study the evolution of the probability density of the ImageNet dataset across the hidden layers in some state-of-the-art DCNs. We find that the initial layers generate a unimodal probability density getting rid of any structure irrelevant for classification. In subsequent layers density peaks arise in a hierarchical fashion that mirrors the semantic hierarchy of the concepts. Density peaks corresponding to single categories appear only close to the output and via a very sharp transition which resembles the nucleation process of a heterogeneous liquid. This process leaves a footprint in the probability density of the output layer where the topography of the peaks allows reconstructing the semantic relationships of the categories.

[Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains](#)

- Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, Ren Ng
- abstract@[open-review](#): We show that passing input points through a simple Fourier feature mapping enables a multilayer perceptron (MLP) to learn high-frequency functions in low-dimensional problem domains. These results shed light on recent advances in computer vision and graphics that achieve state-of-the-art results by using MLPs to represent complex 3D objects and scenes. Using tools from the neural tangent kernel (NTK) literature, we show that a standard MLP has impractically slow convergence to high frequency signal components. To overcome this spectral bias, we use a Fourier feature mapping to transform the effective NTK into a stationary kernel with a tunable bandwidth. We suggest an approach for selecting problem-specific Fourier features that greatly improves the performance of MLPs for low-dimensional regression tasks relevant to the computer vision and graphics communities.

Graph Geometry Interaction Learning

- Shichao Zhu, Shirui Pan, Chuan Zhou, Jia Wu, Yanan Cao, Bin Wang
- abstract@[open-review](#): While numerous approaches have been developed to embed graphs into either Euclidean or hyperbolic spaces, they do not fully utilize the information available in graphs, or lack the flexibility to model intrinsic complex graph geometry. To utilize the strength of both Euclidean and hyperbolic geometries, we develop a novel Geometry Interaction Learning (GIL) method for graphs, a well-suited and efficient alternative for learning abundant geometric properties in graph. GIL captures a more informative internal structural features with low dimensions while maintaining conformal invariance of each space. Furthermore, our method endows each node the freedom to determine the importance of each geometry space via a flexible dual feature interaction learning and probability assembling mechanism. Promising experimental results are presented for five benchmark datasets on node classification and link prediction tasks.

Differentiable Augmentation for Data-Efficient GAN Training

- Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, Song Han
- abstract@[open-review](#): The performance of generative adversarial networks (GANs) heavily deteriorates given a limited amount of training data. This is mainly because the discriminators memorizing the exact training set. To combat it, we propose Differentiable Augmentation (DiffAugment), a simple method that improves the data efficiency of GANs by imposing various types of differentiable augmentations on both real and fake samples. Previous attempts to directly augment the training data manipulate the distribution of real images, yielding little benefit; DiffAugment enables us to adopt the differentiable augmentation for the generated samples, effectively stabilizes training, and leads to better convergence. Experiments demonstrate consistent gains of our method over a variety of GAN architectures and loss functions for both unconditional and class-conditional generation. With DiffAugment, we achieve a state-of-the-art FID of 6.80 with an IS of 100.8 on ImageNet 128×128 and 2-4×2-4 reductions of FID given 1,000 images on FFHQ and LSUN. Furthermore, with only 20% training data, we can match the top performance on CIFAR-10 and CIFAR-100. Finally, our method can generate high-fidelity images using only 100 images without pre-training, while being on par with existing transfer learning algorithms. Code is available at <https://github.com/mit-han-lab/data-efficient-gans>.

Heuristic Domain Adaptation

- Shuhao Cui, Xuan Jin, Shuhui Wang, Yuan He, Qingming Huang
- abstract@[open-review](#): In visual domain adaptation (DA), separating the domain-specific characteristics from the domain-invariant representations is an ill-posed problem. Existing methods apply different kinds of priors or directly minimize the domain discrepancy to address this problem, which lack flexibility in handling real-world situations. Another research pipeline expresses the domain-specific information as a gradual transferring process, which tends to be suboptimal in accurately removing the domain-specific properties. In this paper, we address the modeling of domain-invariant and domain-specific information from the heuristic search perspective. We identify the characteristics in the existing representations that lead to larger domain discrepancy as the heuristic representations. With the guidance of heuristic representations, we formulate a principled framework of Heuristic Domain Adaptation (HDA) with well-founded theoretical guarantees. To perform HDA, the cosine similarity scores and independence measurements between domain-invariant and domain-specific representations are cast into the constraints at the initial and final states during the learning procedure. Similar to the final condition of heuristic search, we further derive a constraint enforcing the final range of heuristic network output to be small. Accordingly, we propose Heuristic Domain Adaptation Network (HDAN), which explicitly learns the domain-invariant and domain-specific representations with the above mentioned constraints. Extensive experiments show that HDAN has exceeded state-of-the-art on unsupervised DA, multi-source DA and semi-supervised DA. The code is available at <https://github.com/cuishuhao/HDA>.

Learning Certified Individually Fair Representations

- Anian Ruoss, Mislav Balunovic, Marc Fischer, Martin Vechev
- abstract@[open-review](#): Fair representation learning provides an effective way of enforcing fairness constraints without compromising utility for downstream users. A desirable family of such fairness constraints, each requiring similar treatment for similar individuals, is known as individual fairness. In this work, we introduce the first method that enables data consumers to obtain certificates of individual fairness for existing and new data points. The key idea is to map similar individuals to close latent representations and leverage this latent proximity to certify individual fairness. That is, our method enables the data producer to learn and certify a representation where for a data point all similar individuals are at l -infinity distance at most epsilon, thus allowing data consumers to certify individual fairness by proving epsilon-robustness of their classifier. Our experimental evaluation on five real-world datasets and several fairness constraints demonstrates the expressivity and scalability of our approach.

Part-dependent Label Noise: Towards Instance-dependent Label Noise

- Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, Masashi Sugiyama
- abstract@[open-review](#): Learning with the \textit{instance-dependent} label noise is challenging, because it is hard to model such real-world noise. Note that there are psychological and physiological evidences showing that we humans perceive instances by decomposing them into parts. Annotators are therefore more likely to annotate instances based on the parts rather than the whole instances, where a wrong mapping from parts to classes may cause the instance-dependent label noise. Motivated by this human cognition, in this paper, we approximate the instance-dependent label noise by exploiting \textit{part-dependent} label noise. Specifically, since instances can be approximately reconstructed by a combination of parts, we approximate the instance-dependent \textit{transition matrix} for an instance by a combination of the transition matrices for the parts of the instance. The transition matrices for parts can be learned by exploiting anchor points (i.e., data points that belong to a specific class almost surely). Empirical evaluations on synthetic and real-world datasets demonstrate our method is superior to the state-of-the-art approaches for learning from the instance-dependent label noise.

Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization

- Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, H. Vincent Poor
- abstract@[open-review](#): In federated learning, heterogeneity in the clients' local datasets and computation speeds results in large variations in the number of local updates performed by each client in each communication round. Naive weighted aggregation of such models causes objective inconsistency, that is, the global model converges to a stationary point of a mismatched objective function which can be arbitrarily different from the true objective. This paper provides a general framework to analyze the convergence of federated heterogeneous optimization algorithms. It subsumes previously proposed methods such as FedAvg and FedProx and provides the first principled understanding of the solution bias and the convergence slowdown due to objective inconsistency. Using insights from this analysis, we propose FedNova, a normalized averaging method that eliminates objective inconsistency while preserving fast error convergence.

An Improved Analysis of (Variance-Reduced) Policy Gradient and Natural Policy Gradient Methods

- Yanli Liu, Kaiqing Zhang, Tamer Basar, Wotao Yin
- abstract@[open-review](#): In this paper, we revisit and improve the convergence of policy gradient (PG), natural PG (NPG) methods, and their variance-reduced variants, under general smooth policy parametrizations. More specifically, with the Fisher information matrix of the policy being positive definite:

i) we show that a state-of-the-art variance-reduced PG method, which has only been shown to converge to stationary points, converges to the globally optimal value up to some inherent function approximation error due to policy parametrization; ii) we show that NPG enjoys a lower sample complexity; iii) we propose SRVR-NPG, which incorporates variance-reduction into the NPG update. Our improvements follow from an observation that the convergence of (variance-reduced) PG and NPG methods can improve each other: the stationary convergence analysis of PG can be applied on NPG as well, and the global convergence analysis of NPG can help to establish the global convergence of (variance-reduced) PG methods. Our analysis carefully integrates the advantages of these two lines of works. Thanks to this improvement, we have also made variance-reduction for NPG possible for the first time, with both global convergence and an efficient finite-sample complexity.

[Geometric Exploration for Online Control](#)

- Orestis Plevrakis, Elad Hazan
- abstract@[open-review](#): We study the control of an \emph{unknown} linear dynamical system under general convex costs. The objective is minimizing regret vs the class of strongly-stable linear policies. In this work, we first consider the case of known cost functions, for which we design the first polynomial-time algorithm with $n^3\sqrt{T}$ -regret, where n is the dimension of the state plus the dimension of control input. The \sqrt{T} -horizon dependence is optimal, and improves upon the previous best known bound of $T^{2/3}$. The main component of our algorithm is a novel geometric exploration strategy: we adaptively construct a sequence of barycentric spanners in an over-parameterized policy space. Second, we consider the case of bandit feedback, for which we give the first polynomial-time algorithm with $\text{poly}(n)\sqrt{T}$ -regret, building on Stochastic Bandit Convex Optimization.

[Automatic Curriculum Learning through Value Disagreement](#)

- Yunzhi Zhang, Pieter Abbeel, Lerrel Pinto
- abstract@[open-review](#): Continually solving new, unsolved tasks is the key to learning diverse behaviors. Through reinforcement learning (RL), we have made massive strides towards solving tasks that have a single goal. However, in the multi-task domain, where an agent needs to reach multiple goals, the choice of training goals can largely affect sample efficiency. When biological agents learn, there is often an organized and meaningful order to which learning happens. Inspired by this, we propose setting up an automatic curriculum for goals that the agent needs to solve. Our key insight is that if we can sample goals at the frontier of the set of goals that an agent is able to reach, it will provide a significantly stronger learning signal compared to randomly sampled goals. To operationalize this idea, we introduce a goal proposal module that prioritizes goals that maximize the epistemic uncertainty of the Q-function of the policy. This simple technique samples goals that are neither too hard nor too easy for the agent to solve, hence enabling continual improvement. We evaluate our method across 13 multi-goal robotic tasks and 5 navigation tasks, and demonstrate performance gains over current state-of-the-art methods.

[MRI Banding Removal via Adversarial Training](#)

- Aaron Defazio, Tullie Murrell, Michael Recht
- abstract@[open-review](#): MR images reconstructed from sub-sampled Cartesian data using deep learning techniques show a characteristic banding (sometimes described as streaking), which is particularly strong in low signal-to-noise regions of the reconstructed image. In this work, we propose the use of an adversarial loss that penalizes banding structures without requiring any human annotation. Our technique greatly reduces the appearance of banding, without requiring any additional computation or post-processing at reconstruction time. We report the results of a blind comparison against a strong baseline by a group of expert evaluators (board-certified radiologists), where our approach is ranked superior at banding removal with no statistically significant loss of detail. A reference implementation of our method is available in the supplementary material.

[The NetHack Learning Environment](#)

- Heinrich Kästner, Nantas Nardelli, Alexander Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, Tim Rocktäschel
- abstract@[open-review](#): Progress in Reinforcement Learning (RL) algorithms goes hand-in-hand with the development of challenging environments that test the limits of current methods. While existing RL environments are either sufficiently complex or based on fast simulation, they are rarely both. Here, we present the NetHack Learning Environment (NLE), a scalable, procedurally generated, stochastic, rich, and challenging environment for RL research based on the popular single-player terminal-based roguelike game, NetHack. We argue that NetHack is sufficiently complex to drive long-term research on problems such as exploration, planning, skill acquisition, and language-conditioned RL, while dramatically reducing the computational resources required to gather a large amount of experience. We compare NLE and its task suite to existing alternatives, and discuss why it is an ideal medium for testing the robustness and systematic generalization of RL agents. We demonstrate empirical success for early stages of the game using a distributed Deep RL baseline and Random Network Distillation exploration, alongside qualitative analysis of various agents trained in the environment. NLE is open source and available at <https://github.com/facebookresearch/nle>.

[Language and Visual Entity Relationship Graph for Agent Navigation](#)

- Yicong Hong, Cristian Rodriguez, Yuankai Qi, Qi Wu, Stephen Gould
- abstract@[open-review](#): Vision-and-Language Navigation (VLN) requires an agent to navigate in a real-world environment following natural language instructions. From both the textual and visual perspectives, we find that the relationships among the scene, its objects, and directional cues are essential for the agent to interpret complex instructions and correctly perceive the environment. To capture and utilize the relationships, we propose a novel Language and Visual Entity Relationship Graph for modelling the inter-modal relationships between text and vision, and the intra-modal relationships among visual entities. We propose a message passing algorithm for propagating information between language elements and visual entities in the graph, which we then combine to determine the next action to take. Experiments show that by taking advantage of the relationships we are able to improve over state-of-the-art. On the Room-to-Room (R2R) benchmark, our method achieves the new best performance on the test unseen split with success rate weighted by path length of 52%. On the Room-for-Room (R4R) dataset, our method significantly improves the previous best from 13% to 34% on the success weighted by normalized dynamic time warping.

[ICAM: Interpretable Classification via Disentangled Representations and Feature Attribution Mapping](#)

- Cher Bass, Mariana da Silva, Carole Sudre, Petru-Daniel Tudor, Stephen Smith, Emma Robinson
- abstract@[open-review](#): Feature attribution (FA), or the assignment of class-relevance to different locations in an image, is important for many classification problems but is particularly crucial within the neuroscience domain, where accurate mechanistic models of behaviours, or disease, require knowledge of all features discriminative of a trait. At the same time, predicting class relevance from brain images is challenging as phenotypes are typically heterogeneous, and changes occur against a background of significant natural variation. Here, we present a novel framework for creating class specific FA maps through image-to-image translation. We propose the use of a VAE-GAN to explicitly disentangle class relevance from background features for improved interpretability properties, which results in meaningful FA maps. We validate our method on 2D and 3D brain image datasets of dementia (ADNI dataset), ageing (UK Biobank), and (simulated) lesion detection. We show that FA maps generated by our method outperform baseline FA methods when validated against ground truth. More significantly, our approach is the first to use latent space sampling to support exploration of phenotype variation.

Spectra of the Conjugate Kernel and Neural Tangent Kernel for linear-width neural networks

- Zhou Fan, Zhichao Wang
- abstract@[open-review](#): We study the eigenvalue distributions of the Conjugate Kernel and Neural Tangent Kernel associated to multi-layer feedforward neural networks. In an asymptotic regime where network width is increasing linearly in sample size, under random initialization of the weights, and for input samples satisfying a notion of approximate pairwise orthogonality, we show that the eigenvalue distributions of the CK and NTK converge to deterministic limits. The limit for the CK is described by iterating the Marenko-Pastur map across the hidden layers. The limit for the NTK is equivalent to that of a linear combination of the CK matrices across layers, and may be described by recursive fixed-point equations that extend this Marenko-Pastur map. We demonstrate the agreement of these asymptotic predictions with the observed spectra for both synthetic and CIFAR-10 training data, and we perform a small simulation to investigate the evolutions of these spectra over training.

No-Regret Learning Dynamics for Extensive-Form Correlated Equilibrium

- Andrea Celli, Alberto Marchesi, Gabriele Farina, Nicola Gatti
- abstract@[open-review](#): The existence of simple, uncoupled no-regret dynamics that converge to correlated equilibria in normal-form games is a celebrated result in the theory of multi-agent systems. Specifically, it has been known for more than 20 years that when all players seek to minimize their internal regret in a repeated normal-form game, the empirical frequency of play converges to a normal-form correlated equilibrium. Extensive-form (that is, tree-form) games generalize normal-form games by modeling both sequential and simultaneous moves, as well as private information. Because of the sequential nature and presence of partial information in the game, extensive-form correlation has significantly different properties than the normal-form counterpart, many of which are still open research directions. Extensive-form correlated equilibrium (EFCE) has been proposed as the natural extensive-form counterpart to normal-form correlated equilibrium. However, it was currently unknown whether EFCE emerges as the result of uncoupled agent dynamics. In this paper, we give the first uncoupled no-regret dynamics that converge to the set of EFCEs in n-player general-sum extensive-form games with perfect recall. First, we introduce a notion of trigger regret in extensive-form games, which extends that of internal regret in normal-form games. When each player has low trigger regret, the empirical frequency of play is a close to an EFCE. Then, we give an efficient no-trigger-regret algorithm. Our algorithm decomposes trigger regret into local subproblems at each decision point for the player, and constructs a global strategy of the player from the local solutions at each decision point.

Estimating weighted areas under the ROC curve

- Andreas Maurer, Massimiliano Pontil
- abstract@[open-review](#): Exponential bounds on the estimation error are given for the plug-in estimator of weighted areas under the ROC curve. The bounds hold for single score functions and uniformly over classes of functions, whose complexity can be controlled by Gaussian or Rademacher averages. The results justify learning algorithms which select score functions to maximize the empirical partial area under the curve (pAUC). They also illustrate the use of some recent advances in the theory of nonlinear empirical processes.

Can Implicit Bias Explain Generalization? Stochastic Convex Optimization as a Case Study

- Assaf Dauber, Meir Feder, Tomer Koren, Roi Livni
- abstract@[open-review](#): The notion of implicit bias, or implicit regularization, has been suggested as a means to explain the surprising generalization ability of modern-days overparameterized learning algorithms. This notion refers to the tendency of the optimization algorithm towards a certain structured solution that often generalizes well. Recently, several papers have studied implicit regularization and were able to identify this phenomenon in various scenarios. We revisit this paradigm in arguably the simplest non-trivial setup, and study the implicit bias of Stochastic Gradient Descent (SGD) in the context of Stochastic Convex Optimization. As a first step, we provide a simple construction that rules out the existence of a \text{distribution-independent} implicit regularizer that governs the generalization ability of SGD. We then demonstrate a learning problem that rules out a very general class of \text{distribution-dependent} implicit regularizers from explaining generalization, which includes strongly convex regularizers as well as non-degenerate norm-based regularizations. Certain aspects of our constructions point out to significant difficulties in providing a comprehensive explanation of an algorithm's generalization performance by solely arguing about its implicit regularization properties.

Generalized Hindsight for Reinforcement Learning

- Alexander Li, Lerrel Pinto, Pieter Abbeel
- abstract@[open-review](#): One of the key reasons for the high sample complexity in reinforcement learning (RL) is the inability to transfer knowledge from one task to another. In standard multi-task RL settings, low-reward data collected while trying to solve one task provides little to no signal for solving that particular task and is hence effectively wasted. However, we argue that this data, which is uninformative for one task, is likely a rich source of information for other tasks. To leverage this insight and efficiently reuse data, we present Generalized Hindsight: an approximate inverse reinforcement learning technique for relabeling behaviors with the right tasks. Intuitively, given a behavior generated under one task, Generalized Hindsight returns a different task that the behavior is better suited for. Then, the behavior is relabeled with this new task before being used by an off-policy RL optimizer. Compared to standard relabeling techniques, Generalized Hindsight provides a substantially more efficient re-use of samples, which we empirically demonstrate on a suite of multi-task navigation and manipulation tasks.

Critic Regularized Regression

- Ziyu Wang, Alexander Novikov, Konrad Zolna, Josh S. Merel, Jost Tobias Springenberg, Scott E. Reed, Bobak Shahriari, Noah Siegel, Caglar Gulcehre, Nicolas Heess, Nando de Freitas
- abstract@[open-review](#): Offline reinforcement learning (RL), also known as batch RL, offers the prospect of policy optimization from large pre-recorded datasets without online environment interaction. It addresses challenges with regard to the cost of data collection and safety, both of which are particularly pertinent to real-world applications of RL. Unfortunately, most off-policy algorithms perform poorly when learning from a fixed dataset. In this paper, we propose a novel offline RL algorithm to learn policies from data using a form of critic-regularized regression (CRR). We find that CRR performs surprisingly well and scales to tasks with high-dimensional state and action spaces -- outperforming several state-of-the-art offline RL algorithms by a significant margin on a wide range of benchmark tasks.

Boosting Adversarial Training with Hypersphere Embedding

- Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Jun Zhu, Hang Su
- abstract@[open-review](#): Adversarial training (AT) is one of the most effective defenses against adversarial attacks for deep learning models. In this work, we advocate incorporating the hypersphere embedding (HE) mechanism into the AT procedure by regularizing the features onto compact manifolds, which constitutes a lightweight yet effective module to blend in the strength of representation learning. Our extensive analyses reveal that AT and HE are well coupled to benefit the robustness of the adversarially trained models from several aspects. We validate the effectiveness and adaptability of HE by embedding it into the popular AT frameworks including PGD-AT, ALP, and TRADES, as well as the FreeAT and FastAT strategies. In the experiments, we evaluate our methods under a wide range of adversarial attacks on the CIFAR-10 and ImageNet datasets, which verifies that integrating HE can consistently enhance the model robustness for each AT framework with little extra computation.

[Beyond Homophily in Graph Neural Networks: Current Limitations and Effective Designs](#)

- Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, Danai Koutra
- abstract@[open-review](#): We investigate the representation power of graph neural networks in the semi-supervised node classification task under heterophily or low homophily, i.e., in networks where connected nodes may have different class labels and dissimilar features. Many popular GNNs fail to generalize to this setting, and are even outperformed by models that ignore the graph structure (e.g., multilayer perceptrons). Motivated by this limitation, we identify a set of key designs—“ego- and neighbor-embedding separation, higher-order neighborhoods, and combination of intermediate representations”—that boost learning from the graph structure under heterophily. We combine them into a graph neural network, H2GCN, which we use as the base method to empirically evaluate the effectiveness of the identified designs. Going beyond the traditional benchmarks with strong homophily, our empirical analysis shows that the identified designs increase the accuracy of GNNs by up to 40% and 27% over models without them on synthetic and real networks with heterophily, respectively, and yield competitive performance under homophily.

[Modeling Continuous Stochastic Processes with Dynamic Normalizing Flows](#)

- Ruizhi Deng, Bo Chang, Marcus A. Brubaker, Greg Mori, Andreas Lehrmann
- abstract@[open-review](#): Normalizing flows transform a simple base distribution into a complex target distribution and have proved to be powerful models for data generation and density estimation. In this work, we propose a novel type of normalizing flow driven by a differential deformation of the continuous-time Wiener process. As a result, we obtain a rich time series model whose observable process inherits many of the appealing properties of its base process, such as efficient computation of likelihoods and marginals. Furthermore, our continuous treatment provides a natural framework for irregular time series with an independent arrival process, including straightforward interpolation. We illustrate the desirable properties of the proposed model on popular stochastic processes and demonstrate its superior flexibility to variational RNN and latent ODE baselines in a series of experiments on synthetic and real-world data.

[Efficient Online Learning of Optimal Rankings: Dimensionality Reduction via Gradient Descent](#)

- Dimitris Fotakis, Thanasis Lianeas, Georgios Piliouras, Stratis Skoulakis
- abstract@[open-review](#): We consider a natural model of online preference aggregation, where sets of preferred items $R_1, R_2, \dots, R_t, \dots$, along with a demand for k_t items in each R_t , appear online. Without prior knowledge of (R_t, k_t) , the learner maintains a ranking π_t aiming that at least k_t items from R_t appear high in π_t . This is a fundamental problem in preference aggregation with applications to e.g., ordering product or news items in web pages based on user scrolling and click patterns. The widely studied Generalized Min-Sum-Set-Cover (GMSSC) problem serves as a formal model for the setting above. GMSSC is NP-hard and the standard application of no-regret online learning algorithms is computationally inefficient, because they operate in the space of rankings. In this work, we show how to achieve low regret for GMSSC in polynomial-time. We employ dimensionality reduction from rankings to the space of doubly stochastic matrices, where we apply Online Gradient Descent. A key step is to show how subgradients can be computed efficiently, by solving the dual of a configuration LP. Using deterministic and randomized rounding schemes, we map doubly stochastic matrices back to rankings with a small loss in the GMSSC objective.

[Training Normalizing Flows with the Information Bottleneck for Competitive Generative Classification](#)

- Lynton Ardizzone, Radek Mackowiak, Carsten Rother, Ullrich Kästner
- abstract@[open-review](#): The Information Bottleneck (IB) objective uses information theory to formulate a task-performance versus robustness trade-off. It has been successfully applied in the standard discriminative classification setting. We pose the question whether the IB can also be used to train generative likelihood models such as normalizing flows. Since normalizing flows use invertible network architectures (INNs), they are information-preserving by construction. This seems contradictory to the idea of a bottleneck. In this work, firstly, we develop the theory and methodology of IB-INNs, a class of conditional normalizing flows where INNs are trained using the IB objective: Introducing a small amount of controlled information loss allows for an asymptotically exact formulation of the IB, while keeping the INN's generative capabilities intact. Secondly, we investigate the properties of these models experimentally, specifically used as generative classifiers. This model class offers advantages such as improved uncertainty quantification and out-of-distribution detection, but traditional generative classifier solutions suffer considerably in classification accuracy. We find the trade-off parameter in the IB controls a mix of generative capabilities and accuracy close to standard classifiers. Empirically, our uncertainty estimates in this mixed regime compare favourably to conventional generative and discriminative classifiers. Code is provided in the supplement.

[Detecting Hands and Recognizing Physical Contact in the Wild](#)

- Supreeth Narasimhaswamy, Trung Nguyen, Minh Hoai Nguyen
- abstract@[open-review](#): We investigate a new problem of detecting hands and recognizing their physical contact state in unconstrained conditions. This is a challenging inference task given the need to reason beyond the local appearance of hands. The lack of training annotations indicating which object or parts of an object the hand is in contact with further complicates the task. We propose a novel convolutional network based on Mask-RCNN that can jointly learn to localize hands and predict their physical contact to address this problem. The network uses outputs from another object detector to obtain locations of objects present in the scene. It uses these outputs and hand locations to recognize the hand's contact state using two attention mechanisms. The first attention mechanism is based on the hand and a region's affinity, enclosing the hand and the object, and densely pools features from this region to the hand region. The second attention module adaptively selects salient features from this plausible region of contact. To develop and evaluate our method's performance, we introduce a large-scale dataset called ContactHands, containing unconstrained images annotated with hand locations and contact states. The proposed network, including the parameters of attention modules, is end-to-end trainable. This network achieves approximately 7% relative improvement over a baseline network that was built on the vanilla Mask-RCNN architecture and trained for recognizing hand contact states.

[On the Theory of Transfer Learning: The Importance of Task Diversity](#)

- Nilesh Tripuraneni, Michael Jordan, Chi Jin
- abstract@[open-review](#): We provide new statistical guarantees for transfer learning via representation learning--when transfer is achieved by learning a feature representation shared across different tasks. This enables learning on new tasks using far less data than is required to learn them in isolation. Formally, we consider $t+1$ tasks parameterized by functions of the form $f_j \circ h$ in a general function class $F \circ H$, where each f_j is a task-specific function in F and h is the shared representation in H . Letting $C(\cdot)$ denote the complexity measure of the function class, we show that for diverse training tasks (1) the sample complexity needed to learn the shared representation across the first t training tasks scales as $C(H) + t C(F)$, despite no explicit access to a signal from the feature representation and (2) with an accurate estimate of the representation, the sample complexity needed to learn a new task scales only with $C(F)$. Our results depend upon a new general notion of task diversity--applicable to models with general tasks, features, and losses--as well as a novel chain rule for Gaussian complexities. Finally, we exhibit the utility of our general framework in several models of importance in the literature.

[Finite-Time Analysis of Round-Robin Kullback-Leibler Upper Confidence Bounds for Optimal Adaptive Allocation with Multiple Plays and Markovian Rewards](#)

- Vrettos Moulos

- abstract@[open-review](#): We study an extension of the classic stochastic multi-armed bandit problem which involves multiple plays and Markovian rewards in the rested bandits setting. In order to tackle this problem we consider an adaptive allocation rule which at each stage combines the information from the sample means of all the arms, with the Kullback-Leibler upper confidence bound of a single arm which is selected in round-robin way. For rewards generated from a one-parameter exponential family of Markov chains, we provide a finite-time upper bound for the regret incurred from this adaptive allocation rule, which reveals the logarithmic dependence of the regret on the time horizon, and which is asymptotically optimal. For our analysis we devise several concentration results for Markov chains, including a maximal inequality for Markov chains, that may be of interest in their own right. As a byproduct of our analysis we also establish asymptotically optimal, finite-time guarantees for the case of multiple plays, and i.i.d. rewards drawn from a one-parameter exponential family of probability densities. Additionally, we provide simulation results that illustrate that calculating Kullback-Leibler upper confidence bounds in a round-robin way, is significantly more efficient than calculating them for every arm at each round, and that the expected regrets of those two approaches behave similarly.

[Neural Star Domain as Primitive Representation](#)

- Yuki Kawana, Yusuke Mukuta, Tatsuya Harada
- abstract@[open-review](#): Reconstructing 3D objects from 2D images is a fundamental task in computer vision. Accurate structured reconstruction by parsimonious and semantic primitive representation further broadens its application. When reconstructing a target shape with multiple primitives, it is preferable that one can instantly access the union of basic properties of the shape such as collective volume and surface, treating the primitives as if they are one single shape. This becomes possible by primitive representation with unified implicit and explicit representations. However, primitive representations in current approaches do not satisfy all of the above requirements at the same time. To solve this problem, we propose a novel primitive representation named neural star domain (NSD) that learns primitive shapes in the star domain. We show that NSD is a universal approximator of the star domain and is not only parsimonious and semantic but also an implicit and explicit shape representation. We demonstrate that our approach outperforms existing methods in image reconstruction tasks, semantic capabilities, and speed and quality of sampling high-resolution meshes.

[Off-Policy Interval Estimation with Lipschitz Value Iteration](#)

- Ziyang Tang, Yihao Feng, Na Zhang, Jian Peng, Qiang Liu
- abstract@[open-review](#): Off-policy evaluation provides an essential tool for evaluating the effects of different policies or treatments using only observed data. When applied to high-stakes scenarios such as medical diagnosis or financial decision-making, it is essential to provide provably correct upper and lower bounds of the expected reward, not just a classical single point estimate, to the end-users, as executing a poor policy can be very costly. In this work, we propose a provably correct method for obtaining interval bounds for off-policy evaluation in a general continuous setting. The idea is to search for the maximum and minimum values of the expected reward among all the Lipschitz Q-functions that are consistent with the observations, which amounts to solving a constrained optimization problem on a Lipschitz function space. We go on to introduce a Lipschitz value iteration method to monotonically tighten the interval, which is simple yet efficient and provably convergent. We demonstrate the practical efficiency of our method on a range of benchmarks.

[Inverse Rational Control with Partially Observable Continuous Nonlinear Dynamics](#)

- Minhae Kwon, Saurabh Daptardar, Paul R. Schrater, Xaq Pitkow
- abstract@[open-review](#): A fundamental question in neuroscience is how the brain creates an internal model of the world to guide actions using sequences of ambiguous sensory information. This is naturally formulated as a reinforcement learning problem under partial observations, where an agent must estimate relevant latent variables in the world from its evidence, anticipate possible future states, and choose actions that optimize total expected reward. This problem can be solved by control theory, which allows us to find the optimal actions for a given system dynamics and objective function. However, animals often appear to behave suboptimally. Why? We hypothesize that animals have their own flawed internal model of the world, and choose actions with the highest expected subjective reward according to that flawed model. We describe this behavior as {it rational} but not optimal. The problem of Inverse Rational Control (IRC) aims to identify which internal model would best explain an agent's actions. Our contribution here generalizes past work on Inverse Rational Control which solved this problem for discrete control in partially observable Markov decision processes. Here we accommodate continuous nonlinear dynamics and continuous actions, and impute sensory observations corrupted by unknown noise that is private to the animal. We first build an optimal Bayesian agent that learns an optimal policy generalized over the entire model space of dynamics and subjective rewards using deep reinforcement learning. Crucially, this allows us to compute a likelihood over models for experimentally observable action trajectories acquired from a suboptimal agent. We then find the model parameters that maximize the likelihood using gradient ascent. Our method successfully recovers the true model of rational agents. This approach provides a foundation for interpreting the behavioral and neural dynamics of animal brains during complex tasks.

[Deep Statistical Solvers](#)

- Balthazar Donon, Zhengying Liu, Wenzhuo LIU, Isabelle Guyon, Antoine Marot, Marc Schoenauer
- abstract@[open-review](#): This paper introduces Deep Statistical Solvers (DSS), a new class of trainable solvers for optimization problems, arising e.g., from system simulations. The key idea is to learn a solver that generalizes to a given distribution of problem instances. This is achieved by directly using as loss the objective function of the problem, as opposed to most previous Machine Learning based approaches, which mimic the solutions attained by an existing solver. Though both types of approaches outperform classical solvers with respect to speed for a given accuracy, a distinctive advantage of DSS is that they can be trained without a training set of sample solutions. Focusing on use cases of systems of interacting and interchangeable entities (e.g. molecular dynamics, power systems, discretized PDEs), the proposed approach is instantiated within a class of Graph Neural Networks. Under sufficient conditions, we prove that the corresponding set of functions contains approximations to any arbitrary precision of the actual solution of the optimization problem. The proposed approach is experimentally validated on large linear problems, demonstrating super-generalisation properties; And on AC power grid simulations, on which the predictions of the trained model have a correlation higher than 99.99% with the outputs of the classical Newton-Raphson method (known for its accuracy), while being 2 to 3 orders of magnitude faster.

[Distributionally Robust Parametric Maximum Likelihood Estimation](#)

- Viet Anh Nguyen, Xuhui Zhang, Jose Blanchet, Angelos Georgiou
- abstract@[open-review](#): We consider the parameter estimation problem of a probabilistic generative model prescribed using a natural exponential family of distributions. For this problem, the typical maximum likelihood estimator usually overfits under limited training sample size, is sensitive to noise and may perform poorly on downstream predictive tasks. To mitigate these issues, we propose a distributionally robust maximum likelihood estimator that minimizes the worst-case expected log-loss uniformly over a parametric Kullback-Leibler ball around a parametric nominal distribution. Leveraging the analytical expression of the Kullback-Leibler divergence between two distributions in the same natural exponential family, we show that the min-max estimation problem is tractable in a broad setting, including the robust training of generalized linear models. Our novel robust estimator also enjoys statistical consistency and delivers promising empirical results in both regression and classification tasks.

[Secretary and Online Matching Problems with Machine Learned Advice](#)

- Antonios Antoniadis, Themis Gouleakis, Pieter Kleer, Pavel Kolev
- abstract@[open-review](#): The classical analysis of online algorithms, due to its worst-case nature, can be quite pessimistic when the input instance at hand is far from worst-case. Often this is not an issue with machine learning approaches, which shine in exploiting patterns in past inputs in order to predict the

future. However, such predictions, although usually accurate, can be arbitrarily poor. Inspired by a recent line of work, we augment three well-known online settings with machine learned predictions about the future, and develop algorithms that take them into account. In particular, we study the following online selection problems: (i) the classical secretary problem, (ii) online bipartite matching and (iii) the graphic matroid secretary problem. Our algorithms still come with a worst-case performance guarantee in the case that predictions are subpar while obtaining an improved competitive ratio (over the best-known classical online algorithm for each problem) when the predictions are sufficiently accurate. For each algorithm, we establish a trade-off between the competitive ratios obtained in the two respective cases.

[Deep Transformation-Invariant Clustering](#)

- Tom Monnier, Thibault Groueix, Mathieu Aubry
- abstract@[open-review](#): Recent advances in image clustering typically focus on learning better deep representations. In contrast, we present an orthogonal approach that does not rely on abstract features but instead learns to predict transformations and performs clustering directly in image space. This learning process naturally fits in the gradient-based training of K-means and Gaussian mixture model, without requiring any additional loss or hyper-parameters. It leads us to two new deep transformation-invariant clustering frameworks, which jointly learn prototypes and transformations. More specifically, we use deep learning modules that enable us to resolve invariance to spatial, color and morphological transformations. Our approach is conceptually simple and comes with several advantages, including the possibility to easily adapt the desired invariance to the task and a strong interpretability of both cluster centers and assignments to clusters. We demonstrate that our novel approach yields competitive and highly promising results on standard image clustering benchmarks. Finally, we showcase its robustness and the advantages of its improved interpretability by visualizing clustering results over real photograph collections.

[Overfitting Can Be Harmless for Basis Pursuit, But Only to a Degree](#)

- Peizhong Ju, Xiaojun Lin, Jia Liu
- abstract@[open-review](#): Recently, there have been significant interests in studying the so-called "double-descent" of the generalization error of linear regression models under the overparameterized and overfitting regime, with the hope that such analysis may provide the first step towards understanding why overparameterized deep neural networks (DNN) still generalize well. However, to date most of these studies focused on the min L2-norm solution that overfits the data. In contrast, in this paper we study the overfitting solution that minimizes the L1-norm, which is known as Basis Pursuit (BP) in the compressed sensing literature. Under a sparse true linear regression model with p i.i.d. Gaussian features, we show that for a large range of p up to a limit that grows exponentially with the number of samples n, with high probability the model error of BP is upper bounded by a value that decreases with p. To the best of our knowledge, this is the first analytical result in the literature establishing the double-descent of overfitting BP for finite n and p. Further, our results reveal significant differences between the double-descent of BP and min L2-norm solutions. Specifically, the double-descent upper-bound of BP is independent of the signal strength, and for high SNR and sparse models the descent-floor of BP can be much lower and wider than that of min L2-norm solutions.

[Improving Generalization in Reinforcement Learning with Mixture Regularization](#)

- KAIXIN WANG, Bingyi Kang, Jie Shao, Jiashi Feng
- abstract@[open-review](#): Deep reinforcement learning (RL) agents trained in a limited set of environments tend to suffer overfitting and fail to generalize to unseen testing environments. To improve their generalizability, data augmentation approaches (e.g. cutout and random convolution) are previously explored to increase the data diversity. However, we find these approaches only locally perturb the observations regardless of the training environments, showing limited effectiveness on enhancing the data diversity and the generalization performance. In this work, we introduce a simple approach, named mixreg, which trains agents on a mixture of observations from different training environments and imposes linearity constraints on the observation interpolations and the supervision (e.g. associated reward) interpolations. Mixreg increases the data diversity more effectively and helps learn smoother policies. We verify its effectiveness on improving generalization by conducting extensive experiments on the large-scale Procgen benchmark. Results show mixreg outperforms the well-established baselines on unseen testing environments by a large margin. Mixreg is simple, effective and general. It can be applied to both policy-based and value-based RL algorithms. Code is available at <https://github.com/kaixin96/mixreg>.

[Pontryagin Differentiable Programming: An End-to-End Learning and Control Framework](#)

- Wanxin Jin, Zhaoran Wang, Zhuoran Yang, Shaoshuai Mou
- abstract@[open-review](#): This paper develops a Pontryagin differentiable programming (PDP) methodology, which establishes a unified framework to solve a broad class of learning and control tasks. The PDP distinguishes from existing methods by two novel techniques: first, we differentiate through Pontryagin's Maximum Principle, and this allows to obtain the analytical derivative of a trajectory with respect to tunable parameters within an optimal control system, enabling end-to-end learning of dynamics, policies, or/and control objective functions; and second, we propose an auxiliary control system in the backward pass of the PDP framework, and the output of this auxiliary control system is the analytical derivative of the original system's trajectory with respect to the parameters, which can be iteratively solved using standard control tools. We investigate three learning modes of the PDP: inverse reinforcement learning, system identification, and control/planning. We demonstrate the capability of the PDP in each learning mode on different high-dimensional systems, including multilink robot arm, 6-DoF maneuvering UAV, and 6-DoF rocket powered landing.

[Learning from Aggregate Observations](#)

- Yivan Zhang, Nontawat Charoenphakdee, Zhenguo Wu, Masashi Sugiyama
- abstract@[open-review](#): We study the problem of learning from aggregate observations where supervision signals are given to sets of instances instead of individual instances, while the goal is still to predict labels of unseen individuals. A well-known example is multiple instance learning (MIL). In this paper, we extend MIL beyond binary classification to other problems such as multiclass classification and regression. We present a general probabilistic framework that accommodates a variety of aggregate observations, e.g., pairwise similarity/triplet comparison for classification and mean/difference/rank observation for regression. Simple maximum likelihood solutions can be applied to various differentiable models such as deep neural networks and gradient boosting machines. Moreover, we develop the concept of consistency up to an equivalence relation to characterize our estimator and show that it has nice convergence properties under mild assumptions. Experiments on three problem settings --- classification via triplet comparison and regression via mean/rank observation indicate the effectiveness of the proposed method.

[The Devil is in the Detail: A Framework for Macroscopic Prediction via Microscopic Models](#)

- Yingxiang Yang, Negar Kiyavash, Le Song, Niao He
- abstract@[open-review](#): Macroscopic data aggregated from microscopic events are pervasive in machine learning, such as country-level COVID-19 infection statistics based on city-level data. Yet, many existing approaches for predicting macroscopic behavior only use aggregated data, leaving a large amount of fine-grained microscopic information unused. In this paper, we propose a principled optimization framework for macroscopic prediction by fitting microscopic models based on conditional stochastic optimization. The framework leverages both macroscopic and microscopic information, and adapts to individual microscopic models involved in the aggregation. In addition, we propose efficient learning algorithms with convergence guarantees. In our experiments, we show that the proposed learning framework clearly outperforms other plug-in supervised learning approaches in real-world applications, including the prediction of daily infections of COVID-19 and medicare claims.

Subgraph Neural Networks

- Emily Alsentzer, Samuel Finlayson, Michelle Li, Marinka Zitnik
- abstract@[open-review](#): Deep learning methods for graphs achieve remarkable performance on many node-level and graph-level prediction tasks. However, despite the proliferation of the methods and their success, prevailing Graph Neural Networks (GNNs) neglect subgraphs, rendering subgraph prediction tasks challenging to tackle in many impactful applications. Further, subgraph prediction tasks present several unique challenges: subgraphs can have non-trivial internal topology, but also carry a notion of position and external connectivity information relative to the underlying graph in which they exist. Here, we introduce SubGNN, a subgraph neural network to learn disentangled subgraph representations. We propose a novel subgraph routing mechanism that propagates neural messages between the subgraph's components and randomly sampled anchor patches from the underlying graph, yielding highly accurate subgraph representations. SubGNN specifies three channels, each designed to capture a distinct aspect of subgraph topology, and we provide empirical evidence that the channels encode their intended properties. We design a series of new synthetic and real-world subgraph datasets. Empirical results for subgraph classification on eight datasets show that SubGNN achieves considerable performance gains, outperforming strong baseline methods, including node-level and graph-level GNNs, by 19.8% over the strongest baseline. SubGNN performs exceptionally well on challenging biomedical datasets, where subgraphs have complex topology and even comprise multiple disconnected components.

Demystifying Orthogonal Monte Carlo and Beyond

- Han Lin, Haoxian Chen, Krzysztof M. Choromanski, Tianyi Zhang, Clement Laroche
- abstract@[open-review](#): Orthogonal Monte Carlo (OMC) is a very effective sampling algorithm imposing structural geometric conditions (orthogonality) on samples for variance reduction. Due to its simplicity and superior performance as compared to its Quasi Monte Carlo counterparts, OMC is used in a wide spectrum of challenging machine learning applications ranging from scalable kernel methods to predictive recurrent neural networks, generative models and reinforcement learning. However theoretical understanding of the method remains very limited. In this paper we shed new light on the theoretical principles behind OMC, applying theory of negatively dependent random variables to obtain several new concentration results. As a corollary, we manage to obtain first uniform convergence results for OMCs and consequently, substantially strengthen best known downstream guarantees for kernel ridge regression via OMCs. We also propose novel extensions of the method leveraging theory of algebraic varieties over finite fields and particle algorithms, called Near-Orthogonal Monte Carlo (NOMC). We show that NOMC is the first algorithm consistently outperforming OMC in applications ranging from kernel methods to approximating distances in probabilistic metric spaces.

Optimal Robustness-Consistency Trade-offs for Learning-Augmented Online Algorithms

- Alexander Wei, Fred Zhang
- abstract@[open-review](#): We study the problem of improving the performance of online algorithms by incorporating machine-learned predictions. The goal is to design algorithms that are both consistent and robust, meaning that the algorithm performs well when predictions are accurate and maintains worst-case guarantees. Such algorithms have been studied in a recent line of works due to Lykouris and Vassilvitskii (ICML '18) and Purohit et al (NeurIPS '18). They provide robustness-consistency trade-offs for a variety of online problems. However, they leave open the question of whether these trade-offs are tight, i.e., to what extent to such trade-offs are necessary. In this paper, we provide the first set of non-trivial lower bounds for competitive analysis using machine-learned predictions. We focus on the classic problems of ski-rental and non-clairvoyant scheduling and provide optimal trade-offs in various settings.

A Scalable Approach for Privacy-Preserving Collaborative Machine Learning

- Jinyun So, Basak Guler, Salman Avestimehr
- abstract@[open-review](#): We consider a collaborative learning scenario in which multiple data-owners wish to jointly train a logistic regression model, while keeping their individual datasets private from the other parties. We propose COPML, a fully-decentralized training framework that achieves scalability and privacy-protection simultaneously. The key idea of COPML is to securely encode the individual datasets to distribute the computation load effectively across many parties and to perform the training computations as well as the model updates in a distributed manner on the securely encoded data. We provide the privacy analysis of COPML and prove its convergence. Furthermore, we experimentally demonstrate that COPML can achieve significant speedup in training over the benchmark protocols. Our protocol provides strong statistical privacy guarantees against colluding parties (adversaries) with unbounded computational power, while achieving up to \$16\times\$ speedup in the training time against the benchmark protocols.

Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search

- Jaehyeon Kim, Sungwon Kim, Jungil Kong, Sungroh Yoon
- abstract@[open-review](#): Recently, text-to-speech (TTS) models such as FastSpeech and ParaNet have been proposed to generate mel-spectrograms from text in parallel. Despite the advantage, the parallel TTS models cannot be trained without guidance from autoregressive TTS models as their external aligners. In this work, we propose Glow-TTS, a flow-based generative model for parallel TTS that does not require any external aligner. By combining the properties of flows and dynamic programming, the proposed model searches for the most probable monotonic alignment between text and the latent representation of speech on its own. We demonstrate that enforcing hard monotonic alignments enables robust TTS, which generalizes to long utterances, and employing generative flows enables fast, diverse, and controllable speech synthesis. Glow-TTS obtains an order-of-magnitude speed-up over the autoregressive model, Tacotron 2, at synthesis with comparable speech quality. We further show that our model can be easily extended to a multi-speaker setting.

Towards Learning Convolutions from Scratch

- Behnam Neyshabur
- abstract@[open-review](#): Convolution is one of the most essential components of modern architectures used in computer vision. As machine learning moves towards reducing the expert bias and learning it from data, a natural next step seems to be learning convolution-like structures from scratch. This, however, has proven elusive. For example, current state-of-the-art architecture search algorithms use convolution as one of the existing modules rather than learning it from data. In an attempt to understand the inductive bias that gives rise to convolutions, we investigate minimum description length as a guiding principle and show that in some settings, it can indeed be indicative of the performance of architectures. To find architectures with small description length, we propose beta-LASSO, a simple variant of LASSO algorithm that, when applied on fully-connected networks for image classification tasks, learns architectures with local connections and achieves state-of-the-art accuracies for training fully-connected networks on CIFAR-10 (84.50%), CIFAR-100 (57.76%) and SVHN (93.84%) bridging the gap between fully-connected and convolutional networks.

Cycle-Contrast for Self-Supervised Video Representation Learning

- Quan Kong, Wengen Wei, Ziwei Deng, Tomoaki Yoshinaga, Tomokazu Murakami
- abstract@[open-review](#): We present Cycle-Contrastive Learning (CCL), a novel self-supervised method for learning video representation. Following a nature that there is a belong and inclusion relation of video and its frames, CCL is designed to find correspondences across frames and videos considering the contrastive representation in their domains respectively. It is different from recent approaches that merely learn correspondences across frames or clips. In our method, the frame and video representations are learned from a single network based on an R3D network, with a shared non-linear

transformation for embedding both frame and video features before the cycle-contrastive loss. We demonstrate that the video representation learned by CCL can be transferred well to downstream tasks of video understanding, outperforming previous methods in nearest neighbour retrieval and action recognition tasks on UCF101, HMDB51 and MMAct.

[Posterior Re-calibration for Imbalanced Datasets](#)

- Junjiao Tian, Yen-Cheng Liu, Nathaniel Glaser, Yen-Chang Hsu, Zsolt Kira
- abstract@[open-review](#): Neural Networks can perform poorly when the training label distribution is heavily imbalanced, as well as when the testing data differs from the training distribution. In order to deal with shift in the testing label distribution, which imbalance causes, we motivate the problem from the perspective of an optimal Bayes classifier and derive a prior rebalancing technique that can be solved through a KL-divergence based optimization. This method allows a flexible post-training hyper-parameter to be efficiently tuned on a validation set and effectively modify the classifier margin to deal with this imbalance. We further combine this method with existing likelihood shift methods, re-interpreting them from the same Bayesian perspective, and demonstrating that our method can deal with both problems in a unified way. The resulting algorithm can be conveniently used on probabilistic classification problems agnostic to underlying architectures. Our results on six different datasets and five different architectures show state of art accuracy, including on large-scale imbalanced datasets such as iNaturalist for classification and Synthia for semantic segmentation. Please see <https://github.com/GT-RIPL/UNO-IC.git> for implementation.

[Novelty Search in Representational Space for Sample Efficient Exploration](#)

- Ruo Yu Tao, Vincent Francois-Lavet, Joelle Pineau
- abstract@[open-review](#): We present a new approach for efficient exploration which leverages a low-dimensional encoding of the environment learned with a combination of model-based and model-free objectives. Our approach uses intrinsic rewards that are based on the distance of nearest neighbors in the low dimensional representational space to gauge novelty. We then leverage these intrinsic rewards for sample-efficient exploration with planning routines in representational space for hard exploration tasks with sparse rewards. One key element of our approach is the use of information theoretic principles to shape our representations in a way so that our novelty reward goes beyond pixel similarity. We test our approach on a number of maze tasks, as well as a control problem and show that our exploration approach is more sample-efficient compared to strong baselines.

[Robust Reinforcement Learning via Adversarial training with Langevin Dynamics](#)

- Parameswaran Kamalaruban, Yu-Ting Huang, Ya-Ping Hsieh, Paul Rolland, Cheng Shi, Volkan Cevher
- abstract@[open-review](#): We introduce a \emph{sampling} perspective to tackle the challenging task of training robust Reinforcement Learning (RL) agents. Leveraging the powerful Stochastic Gradient Langevin Dynamics, we present a novel, scalable two-player RL algorithm, which is a sampling variant of the two-player policy gradient method. Our algorithm consistently outperforms existing baselines, in terms of generalization across different training and testing conditions, on several MuJoCo environments. Our experiments also show that, even for objective functions that entirely ignore potential environmental shifts, our sampling approach remains highly robust in comparison to standard RL algorithms.

[Adversarial Blocking Bandits](#)

- Nicholas Bishop, Hau Chan, Debmalya Mandal, Long Tran-Thanh
- abstract@[open-review](#): We consider a general adversarial multi-armed blocking bandit setting where each played arm can be blocked (unavailable) for some time periods and the reward per arm is given at each time period adversarially without obeying any distribution. The setting models scenarios of allocating scarce limited supplies (e.g., arms) where the supplies replenish and can be reused only after certain time periods. We first show that, in the optimization setting, when the blocking durations and rewards are known in advance, finding an optimal policy (e.g., determining which arm per round) that maximises the cumulative reward is strongly NP-hard, eliminating the possibility of a fully polynomial-time approximation scheme (FPTAS) for the problem unless P = NP. To complement our result, we show that a greedy algorithm that plays the best available arm at each round provides an approximation guarantee that depends on the blocking durations and the path variance of the rewards. In the bandit setting, when the blocking durations and rewards are not known, we design two algorithms, RGA and RGA-META, for the case of bounded duration and path variation. In particular, when the variation budget BT is known in advance, RGA can achieve $O(\sqrt{T(2\tilde{D}+K)B})$ dynamic approximate regret. On the other hand, when B_T is not known, we show that the dynamic approximate regret of RGA-META is at most $O((K+\tilde{D})^{1/4}\tilde{B}^{1/2}T^{3/4})$ where \tilde{B} is the maximal path variation budget within each batch of RGA-META (which is provably in order of $O(\sqrt{T})$). We also prove that if either the variation budget or the maximal blocking duration is unbounded, the approximate regret will be at least $\Theta(T)$. We also show that the regret upper bound of RGA is tight if the blocking durations are bounded above by an order of $O(1)$.

[Online Algorithms for Multi-shop Ski Rental with Machine Learned Advice](#)

- Shufan Wang, Jian Li, Shiqiang Wang
- abstract@[open-review](#): We study the problem of augmenting online algorithms with machine learned (ML) advice. In particular, we consider the \emph{multi-shop ski rental} (MSSR) problem, which is a generalization of the classical ski rental problem. In MSSR, each shop has different prices for buying and renting a pair of skis, and a skier has to make decisions on when and where to buy. We obtain both deterministic and randomized online algorithms with provably improved performance when either a single or multiple ML predictions are used to make decisions. These online algorithms have no knowledge about the quality or the prediction error type of the ML prediction. The performance of these online algorithms are robust to the poor performance of the predictors, but improve with better predictions. Extensive experiments using both synthetic and real world data traces verify our theoretical observations and show better performance against algorithms that purely rely on online decision making.

[Multi-label Contrastive Predictive Coding](#)

- Jiaming Song, Stefano Ermon
- abstract@[open-review](#): Variational mutual information (MI) estimators are widely used in unsupervised representation learning methods such as contrastive predictive coding (CPC). A lower bound on MI can be obtained from a multi-class classification problem, where a critic attempts to distinguish a positive sample drawn from the underlying joint distribution from $(m-1)$ negative samples drawn from a suitable proposal distribution. Using this approach, MI estimates are bounded above by $\log m$, and could thus severely underestimate unless m is very large. To overcome this limitation, we introduce a novel estimator based on a multi-label classification problem, where the critic needs to jointly identify \emph{multiple} positive samples at the same time. We show that using the same amount of negative samples, multi-label CPC is able to exceed the $\log m$ bound, while still being a valid lower bound of mutual information. We demonstrate that the proposed approach is able to lead to better mutual information estimation, gain empirical improvements in unsupervised representation learning, and beat the current state-of-the-art in knowledge distillation over 10 out of 13 tasks.

[Rotation-Invariant Local-to-Global Representation Learning for 3D Point Cloud](#)

- SEOHYUN KIM, JaeYoo Park, Bohyung Han
- abstract@[open-review](#): We propose a local-to-global representation learning algorithm for 3D point cloud data, which is appropriate to handle various geometric transformations, especially rotation, without explicit data augmentation with respect to the transformations. Our model takes advantage of

multi-level abstraction based on graph convolutional neural networks, which constructs a descriptor hierarchy to encode rotation-invariant shape information of an input object in a bottom-up manner. The descriptors in each level are obtained from a neural network based on a graph via stochastic sampling of 3D points, which is effective in making the learned representations robust to the variations of input data. The proposed algorithm presents the state-of-the-art performance on the rotation-augmented 3D object recognition and segmentation benchmarks, and we further analyze its characteristics through comprehensive ablative experiments.

[Learning Invariants through Soft Unification](#)

- Nuri Cingillioglu, Alessandra Russo
- abstract@[open-review](#): Human reasoning involves recognising common underlying principles across many examples. The by-products of such reasoning are invariants that capture patterns such as "if someone went somewhere then they are there", expressed using variables "someone" and "somewhere" instead of mentioning specific people or places. Humans learn what variables are and how to use them at a young age. This paper explores whether machines can also learn and use variables solely from examples without requiring human pre-engineering. We propose Unification Networks, an end-to-end differentiable neural network approach capable of lifting examples into invariants and using those invariants to solve a given task. The core characteristic of our architecture is soft unification between examples that enables the network to generalise parts of the input into variables, thereby learning invariants. We evaluate our approach on five datasets to demonstrate that learning invariants captures patterns in the data and can improve performance over baselines.

[One Solution is Not All You Need: Few-Shot Extrapolation via Structured MaxEnt RL](#)

- Saurabh Kumar, Aviral Kumar, Sergey Levine, Chelsea Finn
- abstract@[open-review](#): While reinforcement learning algorithms can learn effective policies for complex tasks, these policies are often brittle to even minor task variations, especially when variations are not explicitly provided during training. One natural approach to this problem is to train agents with manually specified variation in the training task or environment. However, this may be infeasible in practical situations, either because making perturbations is not possible, or because it is unclear how to choose suitable perturbation strategies without sacrificing performance. The key insight of this work is that learning diverse behaviors for accomplishing a task can directly lead to behavior that generalizes to varying environments, without needing to perform explicit perturbations during training. By identifying multiple solutions for the task in a single environment during training, our approach can generalize to new situations by abandoning solutions that are no longer effective and adopting those that are. We theoretically characterize a robustness set of environments that arises from our algorithm and empirically find that our diversity-driven approach can extrapolate to various changes in the environment and task.

[Variational Bayesian Monte Carlo with Noisy Likelihoods](#)

- Luigi Acerbi
- abstract@[open-review](#): Variational Bayesian Monte Carlo (VPMC) is a recently introduced framework that uses Gaussian process surrogates to perform approximate Bayesian inference in models with black-box, non-cheap likelihoods. In this work, we extend VPMC to deal with noisy log-likelihood evaluations, such as those arising from simulation-based models. We introduce new global' acquisition functions, such as expected information gain (EIG) and variational interquartile range (VIQR), which are robust to noise and can be efficiently evaluated within the VPMC setting. In a novel, challenging, noisy-inference benchmark comprising of a variety of models with real datasets from computational and cognitive neuroscience, VPMC+VIQR achieves state-of-the-art performance in recovering the ground-truth posteriors and model evidence. In particular, our method vastly outperforms local' acquisition functions and other surrogate-based inference methods while keeping a small algorithmic cost. Our benchmark corroborates VPMC as a general-purpose technique for sample-efficient black-box Bayesian inference also with noisy models.

[Finite-Sample Analysis of Contractive Stochastic Approximation Using Smooth Convex Envelopes](#)

- Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, Karthikeyan Shanmugam
- abstract@[open-review](#): Stochastic Approximation (SA) is a popular approach for solving fixed-point equations where the information is corrupted by noise. In this paper, we consider an SA involving a contraction mapping with respect to an arbitrary norm, and show its finite-sample error bounds while using different stepsizes. The idea is to construct a smooth Lyapunov function using the generalized Moreau envelope, and show that the iterates of SA have negative drift with respect to that Lyapunov function. Our result is applicable in Reinforcement Learning (RL). In particular, we use it to establish the first-known convergence rate of the V-trace algorithm for off-policy TD-learning [18]. Importantly, our construction results in only a logarithmic dependence of the convergence bound on the size of the state-space.

[Self-Supervised Generative Adversarial Compression](#)

- Chong Yu, Jeff Pool
- abstract@[open-review](#): Deep learning's success has led to larger and larger models to handle more and more complex tasks; trained models often contain millions of parameters. These large models are compute- and memory-intensive, which makes it a challenge to deploy them with latency, throughput, and storage constraints. Some model compression methods have been successfully applied to image classification and detection or language models, but there has been very little work compressing generative adversarial networks (GANs) performing complex tasks. In this paper, we show that a standard model compression technique, weight pruning and knowledge distillation, cannot be applied to GANs using existing methods. We then develop a self-supervised compression technique which uses the trained discriminator to supervise the training of a compressed generator. We show that this framework has compelling performance to high degrees of sparsity, can be easily applied to new tasks and models, and enables meaningful comparisons between different compression granularities.

[An efficient nonconvex reformulation of stagewise convex optimization problems](#)

- Rudy R. Bunel, Oliver Hinder, Srinadh Bhojanapalli, Krishnamurthy Dvijotham
- abstract@[open-review](#): Convex optimization problems with staged structure appear in several contexts, including optimal control, verification of deep neural networks, and isotonic regression. Off-the-shelf solvers can solve these problems but may scale poorly. We develop a nonconvex reformulation designed to exploit this staged structure. Our reformulation has only simple bound constraints, enabling solution via projected gradient methods and their accelerated variants. The method automatically generates a sequence of primal and dual feasible solutions to the original convex problem, making optimality certification easy. We establish theoretical properties of the nonconvex formulation, showing that it is (almost) free of spurious local minima and has the same global optimum as the convex problem. We modify projected gradient descent to avoid spurious local minimizers so it always converges to the global minimizer. For neural network verification, our approach obtains small duality gaps in only a few gradient steps. Consequently, it can provide tight duality gaps for many large-scale verification problems where both off-the-shelf and specialized solvers struggle.

[From Finite to Countable-Armed Bandits](#)

- Anand Kalvit, Assaf Zeevi

- abstract@[open-review](#): We consider a stochastic bandit problem with countably many arms that belong to a finite set of types, each characterized by a unique mean reward. In addition, there is a fixed distribution over types which sets the proportion of each type in the population of arms. The decision maker is oblivious to the type of any arm and to the aforementioned distribution over types, but perfectly knows the total number of types occurring in the population of arms. We propose a fully adaptive online learning algorithm that achieves $O(\log n)$ distribution-dependent expected cumulative regret after any number of plays n , and show that this order of regret is best possible. The analysis of our algorithm relies on newly discovered concentration and convergence properties of optimism-based policies like UCB in finite-armed bandit problems with zero gap, which may be of independent interest.

[Adversarial Distributional Training for Robust Deep Learning](#)

- Yinpeng Dong, Zhijie Deng, Tianyu Pang, Jun Zhu, Hang Su
- abstract@[open-review](#): Adversarial training (AT) is among the most effective techniques to improve model robustness by augmenting training data with adversarial examples. However, most existing AT methods adopt a specific attack to craft adversarial examples, leading to the unreliable robustness against other unseen attacks. Besides, a single attack algorithm could be insufficient to explore the space of perturbations. In this paper, we introduce adversarial distributional training (ADT), a novel framework for learning robust models. ADT is formulated as a minimax optimization problem, where the inner maximization aims to learn an adversarial distribution to characterize the potential adversarial examples around a natural one under an entropic regularizer, and the outer minimization aims to train robust models by minimizing the expected loss over the worst-case adversarial distributions. Through a theoretical analysis, we develop a general algorithm for solving ADT, and present three approaches for parameterizing the adversarial distributions, ranging from the typical Gaussian distributions to the flexible implicit ones. Empirical results on several benchmarks validate the effectiveness of ADT compared with the state-of-the-art AT methods.

[Meta-Learning Stationary Stochastic Process Prediction with Convolutional Neural Processes](#)

- Andrew Foong, Wessel Bruinsma, Jonathan Gordon, Yann Dubois, James Requeima, Richard Turner
- abstract@[open-review](#): Stationary stochastic processes (SPs) are a key component of many probabilistic models, such as those for off-the-grid spatio-temporal data. They enable the statistical symmetry of underlying physical phenomena to be leveraged, thereby aiding generalization. Prediction in such models can be viewed as a translation equivariant map from observed data sets to predictive SPs, emphasizing the intimate relationship between stationarity and equivariance. Building on this, we propose the Convolutional Neural Process (ConvNP), which endows Neural Processes (NPs) with translation equivariance and extends convolutional conditional NPs to allow for dependencies in the predictive distribution. The latter enables ConvNPs to be deployed in settings which require coherent samples, such as Thompson sampling or conditional image completion. Moreover, we propose a new maximum-likelihood objective to replace the standard ELBO objective in NPs, which conceptually simplifies the framework and empirically improves performance. We demonstrate the strong performance and generalization capabilities of ConvNPs on 1D regression, image completion, and various tasks with real-world spatio-temporal data.

[Theory-Inspired Path-Regularized Differential Network Architecture Search](#)

- Pan Zhou, Caiming Xiong, Richard Socher, Steven Chu Hong Hoi
- abstract@[open-review](#): Despite its high search efficiency, differential architecture search (DARTS) often selects network architectures with dominated skip connections which lead to performance degradation. However, theoretical understandings on this issue remain absent, hindering the development of more advanced methods in a principled way. In this work, we solve this problem by theoretically analyzing the effects of various types of operations, e.g. convolution, skip connection and zero operation, to the network optimization. We prove that the architectures with more skip connections can converge faster than the other candidates, and thus are selected by DARTS. This result, for the first time, theoretically and explicitly reveals the impact of skip connections to fast network optimization and its competitive advantage over other types of operations in DARTS. Then we propose a theory-inspired path-regularized DARTS that consists of two key modules: (i) a differential group-structured sparse binary gate introduced for each operation to avoid unfair competition among operations, and (ii) a path-depth-wise regularization used to incite search exploration for deep architectures that often converge slower than shallow ones as shown in our theory and are not well explored during search. Experimental results on image classification tasks validate its advantages. Codes and models will be released.

[Conic Descent and its Application to Memory-efficient Optimization over Positive Semidefinite Matrices](#)

- John C. Duchi, Oliver Hinder, Andrew Naber, Yinyu Ye
- abstract@[open-review](#): We present an extension of the conditional gradient method to problems whose feasible sets are convex cones. We provide a convergence analysis for the method and for variants with nonconvex objectives, and we extend the analysis to practical cases with effective line search strategies. For the specific case of the positive semidefinite cone, we present a memory-efficient version based on randomized matrix sketches and advocate a heuristic greedy step that greatly improves its practical performance. Numerical results on phase retrieval and matrix completion problems indicate that our method can offer substantial advantages over traditional conditional gradient and Burer-Monteiro approaches.

[Learning the Geometry of Wave-Based Imaging](#)

- Konik Kothari, Maarten de Hoop, Ivan Dokmanić
- abstract@[open-review](#): We propose a general physics-based deep learning architecture for wave-based imaging problems. A key difficulty in imaging problems with a varying background wave speed is that the medium ``bends'' the waves differently depending on their position and direction. This space-bending geometry makes the equivariance to translations of convolutional networks an undesired inductive bias. We build an interpretable neural architecture inspired by Fourier integral operators (FIOs) which approximate the wave physics. FIOs model a wide range of imaging modalities, from seismology and radar to Doppler and ultrasound. We focus on learning the geometry of wave propagation captured by FIOs, which is implicit in the data, via a loss based on optimal transport. The proposed FIONet performs significantly better than the usual baselines on a number of imaging inverse problems, especially in out-of-distribution tests.

[Greedy inference with structure-exploiting lazy maps](#)

- Michael Brennan, Daniele Bigoni, Olivier Zahm, Alessio Spantini, Youssef Marzouk
- abstract@[open-review](#): We propose a framework for solving high-dimensional Bayesian inference problems using \emph{structure-exploiting} low-dimensional transport maps or flows. These maps are confined to a low-dimensional subspace (hence, lazy), and the subspace is identified by minimizing an upper bound on the Kullback--Leibler divergence (hence, structured). Our framework provides a principled way of identifying and exploiting low-dimensional structure in an inference problem. It focuses the expressiveness of a transport map along the directions of most significant discrepancy from the posterior, and can be used to build deep compositions of lazy maps, where low-dimensional projections of the parameters are iteratively transformed to match the posterior. We prove weak convergence of the generated sequence of distributions to the posterior, and we demonstrate the benefits of the framework on challenging inference problems in machine learning and differential equations, using inverse autoregressive flows and polynomial maps as examples of the underlying density estimators.

[Nimble: Lightweight and Parallel GPU Task Scheduling for Deep Learning](#)

- Woosuk Kwon, Gyeong-In Yu, Eunji Jeong, Byung-Gon Chun
- abstract@[open-review](#): Deep learning (DL) frameworks take advantage of GPUs to improve the speed of DL inference and training. Ideally, DL frameworks should be able to fully utilize the computation power of GPUs such that the running time depends on the amount of computation assigned to GPUs. Yet, we observe that in scheduling GPU tasks, existing DL frameworks suffer from inefficiencies such as large scheduling overhead and unnecessary serial execution. To this end, we propose Nimble, a DL execution engine that runs GPU tasks in parallel with minimal scheduling overhead. Nimble introduces a novel technique called ahead-of-time (AoT) scheduling. Here, the scheduling procedure finishes before executing the GPU kernel, thereby removing most of the scheduling overhead during run time. Furthermore, Nimble automatically parallelizes the execution of GPU tasks by exploiting multiple GPU streams in a single GPU. Evaluation on a variety of neural networks shows that compared to PyTorch, Nimble speeds up inference and training by up to 22.34Å— and 3.61Å—, respectively. Moreover, Nimble outperforms state-of-the-art inference systems, TensorRT and TVM, by up to 2.81Å— and 1.70Å—, respectively.

[Finding the Homology of Decision Boundaries with Active Learning](#)

- Weizhi Li, Gautam Dasarathy, Karthikeyan Natesan Ramamurthy, Visar Berisha
- abstract@[open-review](#): Accurately and efficiently characterizing the decision boundary of classifiers is important for problems related to model selection and meta-learning. Inspired by topological data analysis, the characterization of decision boundaries using their homology has recently emerged as a general and powerful tool. In this paper, we propose an active learning algorithm to recover the homology of decision boundaries. Our algorithm sequentially and adaptively selects which samples it requires the labels of. We theoretically analyze the proposed framework and show that the query complexity of our active learning algorithm depends naturally on the intrinsic complexity of the underlying manifold. We demonstrate the effectiveness of our framework in selecting best-performing machine learning models for datasets just using their respective homological summaries. Experiments on several standard datasets show the sample complexity improvement in recovering the homology and demonstrate the practical utility of the framework for model selection.

[Reinforced Molecular Optimization with Neighborhood-Controlled Grammars](#)

- Chencheng Xu, Qiao Liu, Minlie Huang, Tao Jiang
- abstract@[open-review](#): A major challenge in the pharmaceutical industry is to design novel molecules with specific desired properties, especially when the property evaluation is costly. Here, we propose MNCE-RL, a graph convolutional policy network for molecular optimization with molecular neighborhood-controlled embedding grammars through reinforcement learning. We extend the original neighborhood-controlled embedding grammars to make them applicable to molecular graph generation and design an efficient algorithm to infer grammatical production rules from given molecules. The use of grammars guarantees the validity of the generated molecular structures. By transforming molecular graphs to parse trees with the inferred grammars, the molecular structure generation task is modeled as a Markov decision process where a policy gradient strategy is utilized. In a series of experiments, we demonstrate that our approach achieves state-of-the-art performance in a diverse range of molecular optimization tasks and exhibits significant superiority in optimizing molecular properties with a limited number of property evaluations.

[Natural Policy Gradient Primal-Dual Method for Constrained Markov Decision Processes](#)

- Dongsheng Ding, Kaiqing Zhang, Tamer Basar, Mihailo Jovanovic
- abstract@[open-review](#): We study sequential decision-making problems in which each agent aims to maximize the expected total reward while satisfying a constraint on the expected total utility. We employ the natural policy gradient method to solve the discounted infinite-horizon Constrained Markov Decision Processes (CMDPs) problem. Specifically, we propose a new Natural Policy Gradient Primal-Dual (NPG-PD) method for CMDPs which updates the primal variable via natural policy gradient ascent and the dual variable via projected sub-gradient descent. Even though the underlying maximization involves a nonconcave objective function and a nonconvex constraint set under the softmax policy parametrization, we prove that our method achieves global convergence with sublinear rates regarding both the optimality gap and the constraint violation. Such a convergence is independent of the size of the state-action space, i.e., it is~dimension-free. Furthermore, for the general smooth policy class, we establish sublinear rates of convergence regarding both the optimality gap and the constraint violation, up to a function approximation error caused by restricted policy parametrization. Finally, we show that two sample-based NPG-PD algorithms inherit such non-asymptotic convergence properties and provide finite-sample complexity guarantees. To the best of our knowledge, our work is the first to establish non-asymptotic convergence guarantees of policy-based primal-dual methods for solving infinite-horizon discounted CMDPs. We also provide computational results to demonstrate merits of our approach.

[Classification Under Misspecification: Halfspaces, Generalized Linear Models, and Evolvability](#)

- Sitan Chen, Frederic Koehler, Ankur Moitra, Morris Yau
- abstract@[open-review](#): In this paper, we revisit the problem of distribution-independently learning halfspaces under Massart noise with rate $\$\\eta\$$. Recent work resolved a long-standing problem in this model of efficiently learning to error $\$\\eta + \\epsilon\$$ for any $\$\\epsilon > 0\$$, by giving an improper learner that partitions space into $\$\\text{poly}(d, 1/\\epsilon)\$$ regions. Here we give a much simpler algorithm and settle a number of outstanding open questions:

(1) We give the first *proper* learner for Massart halfspaces that achieves $\$\\eta + \\epsilon\$$. (2) Based on (1), we develop a blackbox knowledge distillation procedure to convert an arbitrarily complex classifier to an equally good proper classifier. (3) By leveraging a simple but overlooked connection to *evolvability*, we show any SQ algorithm requires super-polynomially many queries to achieve $\$\\mathsf{OPT} + \\epsilon\$$.

We then zoom out to study generalized linear models and give an efficient algorithm for learning under a challenging new corruption model generalizing Massart noise. Finally we study our algorithm for learning halfspaces under Massart noise empirically and find that it exhibits some appealing fairness properties as a byproduct of its strong provable robustness guarantees.

[Certified Defense to Image Transformations via Randomized Smoothing](#)

- Marc Fischer, Maximilian Baader, Martin Vechev
- abstract@[open-review](#): We extend randomized smoothing to cover parameterized transformations (e.g., rotations, translations) and certify robustness in the parameter space (e.g., rotation angle). This is particularly challenging as interpolation and rounding effects mean that image transformations do not compose, in turn preventing direct certification of the perturbed image (unlike certification with $\$\\ell^p\$$ norms). We address this challenge by introducing three different defenses, each with a different guarantee (heuristic, distributional and individual) stemming from the method used to bound the interpolation error. Importantly, in the individual case, we show how to efficiently compute the inverse of an image transformation, enabling us to provide individual guarantees in the online setting. We provide an implementation of all methods at <https://github.com/eth-sri/transformation-smoothing>.

[Estimation of Skill Distribution from a Tournament](#)

- Ali Jadbabaie, Anuran Makur, Devavrat Shah
- abstract@[open-review](#): In this paper, we study the problem of learning the skill distribution of a population of agents from observations of pairwise games in a tournament. These games are played among randomly drawn agents from the population. The agents in our model can be individuals, sports teams, or Wall Street fund managers. Formally, we postulate that the likelihoods of outcomes of games are governed by the parametric Bradley-Terry-Luce (or multinomial logit) model, where the probability of an agent beating another is the ratio between its skill level and the pairwise sum of skill levels, and the

skill parameters are drawn from an unknown, non-parametric skill density of interest. The problem is, in essence, to learn a distribution from noisy, quantized observations. We propose a surprisingly simple and tractable algorithm that learns the skill density with near-optimal minimax mean squared error scaling as $\$n^{-\{1+\backslash varepsilon\}}$, for any $\backslash varepsilon > 0$ $, so long as the density is smooth. Our approach brings together prior work on learning skill parameters from pairwise comparisons with kernel density estimation from non-parametric statistics. Furthermore, we prove information theoretic lower bounds which establish minimax optimality of the skill parameter estimation technique used in our algorithm. These bounds utilize a continuum version of Fano's method along with a careful covering argument. We apply our algorithm to various soccer leagues and world cups, cricket world cups, and mutual funds. We find that the entropy of a learnt distribution provides a quantitative measure of skill, which in turn provides rigorous explanations for popular beliefs about perceived qualities of sporting events, e.g., soccer league rankings. Finally, we apply our method to assess the skill distributions of mutual funds. Our results shed light on the abundance of low quality funds prior to the Great Recession of 2008, and the domination of the industry by more skilled funds after the financial crisis.$

[Reparameterizing Mirror Descent as Gradient Descent](#)

- Ehsan Amid, Manfred K. K. Warmuth
- abstract@[open-review](#): Most of the recent successful applications of neural networks have been based on training with gradient descent updates. However, for some small networks, other mirror descent updates learn provably more efficiently when the target is sparse. We present a general framework for casting a mirror descent update as a gradient descent update on a different set of parameters. In some cases, the mirror descent reparameterization can be described as training a modified network with standard backpropagation. The reparameterization framework is versatile and covers a wide range of mirror descent updates, even cases where the domain is constrained. Our construction for the reparameterization argument is done for the continuous versions of the updates. Finding general criteria for the discrete versions to closely track their continuous counterparts remains an interesting open problem.

[General Control Functions for Causal Effect Estimation from IVs](#)

- Aahlad Puli, Rajesh Ranganath
- abstract@[open-review](#): Causal effect estimation relies on separating the variation in the outcome into parts due to the treatment and due to the confounders. To achieve this separation, practitioners often use external sources of randomness that only influence the treatment called instrumental variables (IVs). We study variables constructed from treatment and IV that help estimate effects, called control functions. We characterize general control functions for effect estimation in a meta-identification result. Then, we show that structural assumptions on the treatment process allow the construction of general control functions, thereby guaranteeing identification. To construct general control functions and estimate effects, we develop the general control function method (GCFN). GCFN's first stage called variational decoupling (VDE) constructs general control functions by recovering the residual variation in the treatment given the IV. Using VDE's control function, GCFN's second stage estimates effects via regression. Further, we develop semi-supervised GCFN to construct general control functions using subsets of data that have both IV and confounders observed as supervision; this needs no structural treatment process assumptions. We evaluate GCFN on low and high dimensional simulated data and on recovering the causal effect of slave export on modern community trust [30]

[Optimal Algorithms for Stochastic Multi-Armed Bandits with Heavy Tailed Rewards](#)

- Kyungjae Lee, Hongjun Yang, Sungbin Lim, Songhwai Oh
- abstract@[open-review](#): In this paper, we consider stochastic multi-armed bandits (MABs) with heavy-tailed rewards, whose p-th moment is bounded by a constant nu_p for 1

[Certified Robustness of Graph Convolution Networks for Graph Classification under Topological Attacks](#)

- Hongwei Jin, Zhan Shi, Venkata Jaya Shankar Ashish Peruri, Xinhua Zhang
- abstract@[open-review](#): Graph convolution networks (GCNs) have become effective models for graph classification. Similar to many deep networks, GCNs are vulnerable to adversarial attacks on graph topology and node attributes. Recently, a number of effective attack and defense algorithms have been designed, but no certificate of robustness has been developed for GCN-based graph classification under topological perturbations with both local and global budgets. In this paper, we propose the first certificate for this problem. Our method is based on Lagrange dualization and convex envelope, which result in tight approximation bounds that are efficiently computable by dynamic programming. When used in conjunction with robust training, it allows an increased number of graphs to be certified as robust.

[Zero-Resource Knowledge-Grounded Dialogue Generation](#)

- Linxiao Li, Can Xu, Wei Wu, YUFAN ZHAO, Xueliang Zhao, Chongyang Tao
- abstract@[open-review](#): While neural conversation models have shown great potentials towards generating informative and engaging responses via introducing external knowledge, learning such a model often requires knowledge-grounded dialogues that are difficult to obtain. To overcome the data challenge and reduce the cost of building a knowledge-grounded dialogue system, we explore the problem under a zero-resource setting by assuming no context-knowledge-response triples are needed for training. To this end, we propose representing the knowledge that bridges a context and a response and the way that the knowledge is expressed as latent variables, and devise a variational approach that can effectively estimate a generation model from independent dialogue corpora and knowledge corpora. Evaluation results on three benchmarks of knowledge-grounded dialogue generation indicate that our model can achieve comparable performance with state-of-the-art methods that rely on knowledge-grounded dialogues for training, and exhibits a good generalization ability over different datasets.

[Targeted Adversarial Perturbations for Monocular Depth Prediction](#)

- Alex Wong, Safa Cicek, Stefano Soatto
- abstract@[open-review](#): We study the effect of adversarial perturbations on the task of monocular depth prediction. Specifically, we explore the ability of small, imperceptible additive perturbations to selectively alter the perceived geometry of the scene. We show that such perturbations can not only globally re-scale the predicted distances from the camera, but also alter the prediction to match a different target scene. We also show that, when given semantic or instance information, perturbations can fool the network to alter the depth of specific categories or instances in the scene, and even remove them while preserving the rest of the scene. To understand the effect of targeted perturbations, we conduct experiments on state-of-the-art monocular depth prediction methods. Our experiments reveal vulnerabilities in monocular depth prediction networks, and shed light on the biases and context learned by them.

[Beyond the Mean-Field: Structured Deep Gaussian Processes Improve the Predictive Uncertainties](#)

- Jakob Lindinger, David Reeb, Christoph Lippert, Barbara Rakitsch
- abstract@[open-review](#): Deep Gaussian Processes learn probabilistic data representations for supervised learning by cascading multiple Gaussian Processes. While this model family promises flexible predictive distributions, exact inference is not tractable. Approximate inference techniques trade off the ability to closely resemble the posterior distribution against speed of convergence and computational efficiency. We propose a novel Gaussian variational family that allows for retaining covariances between latent processes while achieving fast convergence by marginalising out all global latent variables. After providing a proof of how this marginalisation can be done for general covariances, we restrict them to the ones we empirically found to be

most important in order to also achieve computational efficiency. We provide an efficient implementation of our new approach and apply it to several benchmark datasets. It yields excellent results and strikes a better balance between accuracy and calibrated uncertainty estimates than its state-of-the-art alternatives.

[Offline Imitation Learning with a Misspecified Simulator](#)

- Shengyi Jiang, Jingcheng Pang, Yang Yu
- abstract@[open-review](#): In real-world decision-making tasks, learning an optimal policy without a trial-and-error process is an appealing challenge. When expert demonstrations are available, imitation learning that mimics expert actions can learn a good policy efficiently. Learning in simulators is another commonly adopted approach to avoid real-world trials-and-errors. However, neither sufficient expert demonstrations nor high-fidelity simulators are easy to obtain. In this work, we investigate policy learning in the condition of a few expert demonstrations and a simulator with misspecified dynamics. Under a mild assumption that local states shall still be partially aligned under a dynamics mismatch, we propose imitation learning with horizon-adaptive inverse dynamics (HIDIL) that matches the simulator states with expert states in a H -step horizon and accurately recovers actions based on inverse dynamics policies. In the real environment, HIDIL can effectively derive adapted actions from the matched states. Experiments are conducted in four MuJoCo locomotion environments with modified friction, gravity, and density configurations. Experiment results show that HIDIL achieves significant improvement in terms of performance and stability in all of the real environments, compared with imitation learning methods and transferring methods in reinforcement learning.

[Multi-Fidelity Bayesian Optimization via Deep Neural Networks](#)

- Shibo Li, Wei Xing, Robert Kirby, Shandian Zhe
- abstract@[open-review](#): Bayesian optimization (BO) is a popular framework for optimizing black-box functions. In many applications, the objective function can be evaluated at multiple fidelities to enable a trade-off between the cost and accuracy. To reduce the optimization cost, many multi-fidelity BO methods have been proposed. Despite their success, these methods either ignore or over-simplify the strong, complex correlations across the fidelities. While the acquisition function is therefore easy and convenient to calculate, these methods can be inefficient in estimating the objective function. To address this issue, we propose Deep Neural Network Multi-Fidelity Bayesian Optimization (DNN-MFBO) that can flexibly capture all kinds of complicated relationships between the fidelities to improve the objective function estimation and hence the optimization performance. We use sequential, fidelity-wise Gauss-Hermite quadrature and moment-matching to compute a mutual information-based acquisition function in a tractable and highly efficient way. We show the advantages of our method in both synthetic benchmark datasets and real-world applications in engineering design.

[PlanGAN: Model-based Planning With Sparse Rewards and Multiple Goals](#)

- Henry Charlesworth, Giovanni Montana
- abstract@[open-review](#): Learning with sparse rewards remains a significant challenge in reinforcement learning (RL), especially when the aim is to train a policy capable of achieving multiple different goals. To date, the most successful approaches for dealing with multi-goal, sparse reward environments have been model-free RL algorithms. In this work we propose PlanGAN, a model-based algorithm specifically designed for solving multi-goal tasks in environments with sparse rewards. Our method builds on the fact that any trajectory of experience collected by an agent contains useful information about how to achieve the goals observed during that trajectory. We use this to train an ensemble of conditional generative models (GANs) to generate plausible trajectories that lead the agent from its current state towards a specified goal. We then combine these imagined trajectories into a novel planning algorithm in order to achieve the desired goal as efficiently as possible. The performance of PlanGAN has been tested on a number of robotic navigation/manipulation tasks in comparison with a range of model-free reinforcement learning baselines, including Hindsight Experience Replay. Our studies indicate that PlanGAN can achieve comparable performance whilst being around 4-8 times more sample efficient.

[Bad Global Minima Exist and SGD Can Reach Them](#)

- Shengchao Liu, Dimitris Papailiopoulos, Dimitris Achlioptas
- abstract@[open-review](#): Several works have aimed to explain why overparameterized neural networks generalize well when trained by Stochastic Gradient Descent (SGD). The consensus explanation that has emerged credits the randomized nature of SGD for the bias of the training process towards low-complexity models and, thus, for implicit regularization. We take a careful look at this explanation in the context of image classification with common deep neural network architectures. We find that if we do not regularize \emph{explicitly}, then SGD can be easily made to converge to poorly-generalizing, high-complexity models: all it takes is to first train on a random labeling on the data, before switching to properly training with the correct labels. In contrast, we find that in the presence of explicit regularization, pretraining with random labels has no detrimental effect on SGD. We believe that our results give evidence that explicit regularization plays a far more important role in the success of overparameterized neural networks than what has been understood until now. Specifically, in suppressing complicated models that got lucky with the training data, regularization not only makes simple models that fit the data well the global optima, but it also clears the way to make them discoverable by local methods, such as SGD.

[Optimal Prediction of the Number of Unseen Species with Multiplicity](#)

- Yi Hao, Ping Li
- abstract@[open-review](#): Based on a sample of size n , we consider estimating the number of symbols that appear at least μ times in an independent sample of size $a \cdot n$, where a is a given parameter. This formulation includes, as a special case, the well-known problem of inferring the number of unseen species introduced by [Fisher et al.] in 1943 and considered by many others. Of considerable interest in this line of works is the largest a for which the quantity can be accurately predicted. We completely resolve this problem by determining the limit of estimation to be $a \approx (\log n)^{\mu}$, with both lower and upper bounds matching up to constant factors. For the particular case of $\mu = 1$, this implies the recent result by [Orlitsky et al.] on the unseen species problem. Experimental evaluations show that the proposed estimator performs exceptionally well in practice. Furthermore, the estimator is a simple linear combination of symbols' empirical counts, and hence linear-time computable.

[Characterizing Optimal Mixed Policies: Where to Intervene and What to Observe](#)

- Sanghack Lee, Elias Bareinboim
- abstract@[open-review](#): Intelligent agents are continuously faced with the challenge of optimizing a policy based on what they can observe (see) and which actions they can take (do) in the environment where they are deployed. Most policy can be parametrized in terms of these two dimensions, i.e., as a function of what can be seen and done given a certain situation, which we call a \textit{mixed policy}. In this paper, we investigate several properties of the class of mixed policies and provide an efficient and effective characterization, including optimality and non-redundancy. Specifically, we introduce a graphical criterion to identify unnecessary contexts for a set of actions, leading to a natural characterization of non-redundancy of mixed policies. We then derive sufficient conditions under which one strategy can dominate the other with respect to their maximum achievable expected rewards (optimality). This characterization leads to a fundamental understanding of the space of mixed policies and a possible refinement of the agent's strategy so that it converges to the optimum faster and more robustly. One surprising result of the causal characterization is that the agent following a more standard approach --- intervening on all intervenable variables and observing all available contexts --- may be hurting itself, and will never achieve an optimal performance.

[Factor Graph Neural Networks](#)

- Zhen Zhang, Fan Wu, Wee Sun Lee
- abstract@[open-review](#): Most of the successful deep neural network architectures are structured, often consisting of elements like convolutional neural networks and gated recurrent neural networks. Recently, graph neural networks (GNNs) have been successfully applied to graph-structured data such as point cloud and molecular data. These networks often only consider pairwise dependencies, as they operate on a graph structure. We generalize the GNN into a factor graph neural network (FGNN) providing a simple way to incorporate dependencies among multiple variables. We show that FGNN is able to represent Max-Product belief propagation, an approximate inference method on probabilistic graphical models, providing a theoretical understanding on the capabilities of FGNN and related GNNs. Experiments on synthetic and real datasets demonstrate the potential of the proposed architecture.

[A Closer Look at Accuracy vs. Robustness](#)

- Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R. Salakhutdinov, Kamalika Chaudhuri
- abstract@[open-review](#): Current methods for training robust networks lead to a drop in test accuracy, which has led prior works to posit that a robustness-accuracy tradeoff may be inevitable in deep learning. We take a closer look at this phenomenon and first show that real image datasets are actually separated. With this property in mind, we then prove that robustness and accuracy should both be achievable for benchmark datasets through locally Lipschitz functions, and hence, there should be no inherent tradeoff between robustness and accuracy. Through extensive experiments with robustness methods, we argue that the gap between theory and practice arises from two limitations of current methods: either they fail to impose local Lipschitzness or they are insufficiently generalized. We explore combining dropout with robust training methods and obtain better generalization. We conclude that achieving robustness and accuracy in practice may require using methods that impose local Lipschitzness and augmenting them with deep learning generalization techniques.

[Curriculum Learning by Dynamic Instance Hardness](#)

- Tianyi Zhou, Shengjie Wang, Jeffrey Bilmes
- abstract@[open-review](#): A good teacher can adjust the curriculum based on students' learning history. By analogy, in this paper, we study the dynamics of a deep neural network's (DNN) performance on individual samples during its learning process. The observed properties allow us to develop an adaptive curriculum that leads to faster learning of more accurate models. We introduce dynamic instance hardness (DIH), the exponential moving average of a sample's instantaneous hardness (e.g., a loss, or a change in outputs) over the training history. A low DIH indicates that a model retains knowledge about a sample over time, and implies a flat loss landscape for that sample. Moreover, for DNNs, we find that a sample's DIH early in training predicts its DIH in later stages. Hence, we can train a model using samples with higher DIH and safely ignore those with lower DIH. This motivates a DIH guided curriculum learning (DIHCL). Compared to existing CL methods: (1) DIH is more stable over time than using only instantaneous hardness, which is noisy due to stochastic training and DNN's non-smoothness; (2) DIHCL is computationally inexpensive since it uses only a byproduct of back-propagation and thus does not require extra inference. On 11 datasets, DIHCL significantly outperforms random mini-batch SGD and recent CL methods in terms of efficiency and final performance.

[Spin-Weighted Spherical CNNs](#)

- Carlos Esteves, Ameesh Makadia, Kostas Daniilidis
- abstract@[open-review](#): Learning equivariant representations is a promising way to reduce sample and model complexity and improve the generalization performance of deep neural networks. The spherical CNNs are successful examples, producing SO(3)-equivariant representations of spherical inputs. There are two main types of spherical CNNs. The first type lifts the inputs to functions on the rotation group SO(3) and applies convolutions on the group, which are computationally expensive since SO(3) has one extra dimension. The second type applies convolutions directly on the sphere, which are limited to zonal (isotropic) filters, and thus have limited expressivity. In this paper, we present a new type of spherical CNN that allows anisotropic filters in an efficient way, without ever leaving the spherical domain. The key idea is to consider spin-weighted spherical functions, which were introduced in physics in the study of gravitational waves. These are complex-valued functions on the sphere whose phases change upon rotation. We define a convolution between spin-weighted functions and build a CNN based on it. The spin-weighted functions can also be interpreted as spherical vector fields, allowing applications to tasks where the inputs or outputs are vector fields. Experiments show that our method outperforms previous methods on tasks like classification of spherical images, classification of 3D shapes and semantic segmentation of spherical panoramas.

[Learning to Execute Programs with Instruction Pointer Attention Graph Neural Networks](#)

- David Bieber, Charles Sutton, Hugo Larochelle, Daniel Tarlow
- abstract@[open-review](#): Graph neural networks (GNNs) have emerged as a powerful tool for learning software engineering tasks including code completion, bug finding, and program repair. They benefit from leveraging program structure like control flow graphs, but they are not well-suited to tasks like program execution that require far more sequential reasoning steps than number of GNN propagation steps. Recurrent neural networks (RNNs), on the other hand, are well-suited to long sequential chains of reasoning, but they do not naturally incorporate program structure and generally perform worse on the above tasks. Our aim is to achieve the best of both worlds, and we do so by introducing a novel GNN architecture, the Instruction Pointer Attention Graph Neural Networks (IPA-GNN), which achieves improved systematic generalization on the task of learning to execute programs using control flow graphs. The model arises by considering RNNs operating on program traces with branch decisions as latent variables. The IPA-GNN can be seen either as a continuous relaxation of the RNN model or as a GNN variant more tailored to execution. To test the models, we propose evaluating systematic generalization on learning to execute using control flow graphs, which tests sequential reasoning and use of program structure. More practically, we evaluate these models on the task of learning to execute partial programs, as might arise if using the model as a heuristic function in program synthesis. Results show that the IPA-GNN outperforms a variety of RNN and GNN baselines on both tasks.

[AutoPrivacy: Automated Layer-wise Parameter Selection for Secure Neural Network Inference](#)

- Qian Lou, Song Bian, Lei Jiang
- abstract@[open-review](#): Hybrid Privacy-Preserving Neural Network (HPPNN) implementing linear layers by Homomorphic Encryption (HE) and nonlinear layers by Garbled Circuit (GC) is one of the most promising secure solutions to emerging Machine Learning as a Service (MLaaS). Unfortunately, a HPPNN suffers from long inference latency, e.g., ~ 100 seconds per image, which makes MLaaS unsatisfactory. Because HE-based linear layers of a HPPNN cost $\$93 \pm 70\%$ inference latency, it is critical to select a set of HE parameters to minimize computational overhead of linear layers. Prior HPPNNs over-pessimistically select huge HE parameters to maintain large noise budgets, since they use the same set of HE parameters for an entire network and ignore the error tolerance capability of a network.

In this paper, for fast and accurate secure neural network inference, we propose an automated layer-wise parameter selector, AutoPrivacy, that leverages deep reinforcement learning to automatically determine a set of HE parameters for each linear layer in a HPPNN. The learning-based HE parameter selection policy outperforms conventional rule-based HE parameter selection policy. Compared to prior HPPNNs, AutoPrivacy-optimized HPPNNs reduce inference latency by $\sim 53\% \pm 70\%$ with negligible loss of accuracy.

[Baxter Permutation Process](#)

- Masahiro Nakano, Akisato Kimura, Takeshi Yamada, Naonori Ueda
- abstract@[open-review](#): In this paper, a Bayesian nonparametric (BNP) model for Baxter permutations (BPs), termed BP process (BPP) is proposed and applied to relational data analysis. The BPs are a well-studied class of permutations, and it has been demonstrated that there is one-to-one correspondence between BPs and several interesting objects including floorplan partitioning (FP), which constitutes a subset of rectangular partitioning (RP). Accordingly, the BPP can be used as an FP model. We combine the BPP with a multi-dimensional extension of the stick-breaking process called the \it block-breaking process\rm to fill the gap between FP and RP, and obtain a stochastic process on arbitrary RPs. Compared with conventional BNP models for arbitrary RPs, the proposed model is simpler and has a high affinity with Bayesian inference.

[Characterizing emergent representations in a space of candidate learning rules for deep networks](#)

- Yinan Cao, Christopher Summerfield, Andrew Saxe
- abstract@[open-review](#): How are sensory representations learned via experience? Deep learning offers a theoretical toolkit for studying how neural codes emerge under different learning rules. Studies suggesting that representations in deep networks resemble those in biological brains have mostly relied on one specific learning rule: gradient descent, the workhorse behind modern deep learning. However, it remains unclear how robust these emergent representations in deep networks are to this specific choice of learning algorithm. Here we present a continuous two-dimensional space of candidate learning rules, parameterized by levels of top-down feedback and Hebbian learning. We show that this space contains five important candidate learning algorithms as specific points--Gradient Descent, Contrastive Hebbian, quasi-Predictive Coding, Hebbian & Anti-Hebbian. Next, we exhaustively characterize the properties of each rule during learning about hierarchically structured data, and identify zones within this space where deep networks exhibit qualitative signatures of biological learning. We find that while a large set of algorithms achieve zero training error at convergence, only a subset show hallmarks of human semantic development like progressive differentiation and illusory correlations. Further, only a subset adjust intermediate neural representations toward task-relevant representations, indicative of backpropagation-like behavior. Finally, we show that algorithms can dramatically differ in their learned neural representations and dynamics, providing experimentally testable hallmarks of different learning principles. Our findings provide a framework linking diverse neural representational geometries to learning principles which can guide future experiments, and offer evidence about the learning rules likely to be at work in biology.

[Fast, Accurate, and Simple Models for Tabular Data via Augmented Distillation](#)

- Rasool Fakoor, Jonas W. Mueller, Nick Erickson, Pratik Chaudhari, Alexander J. Smola
- abstract@[open-review](#): Automated machine learning (AutoML) can produce complex model ensembles by stacking, bagging, and boosting many individual models like trees, deep networks, and nearest neighbor estimators. While highly accurate, the resulting predictors are large, slow, and opaque as compared to their constituents. To improve the deployment of AutoML on tabular data, we propose FAST-DAD to distill arbitrarily-complex ensemble predictors into individual models like boosted trees, random forests, and deep networks. At the heart of our approach is a data augmentation strategy based on Gibbs sampling from a self-attention pseudolikelihood estimator. Across 30 datasets spanning regression and binary/multiclass classification tasks, FAST-DAD distillation produces significantly better individual models than one obtains through standard training on the original data. Our individual distilled models are over 10x faster and more accurate than ensemble predictors produced by AutoML tools like H2O/AutoSklearn.

[Adaptive Probing Policies for Shortest Path Routing](#)

- Aditya Bhaskara, Sreenivas Gollapudi, Kostas Kollias, Kamesh Munagala
- abstract@[open-review](#): Inspired by traffic routing applications, we consider the problem of finding the shortest path from a source $\$s\$$ to a destination $\$t\$$ in a graph, when the lengths of the edges are unknown. Instead, we are given \{em hints\} or predictions of the edge lengths from a collection of ML models, trained possibly on historical data and other contexts in the network. Additionally, we assume that the true length of any candidate path can be obtained by \{em probing\} an up-to-date snapshot of the network. However, each probe introduces a latency, and thus the goal is to minimize the number of probes while finding a near-optimal path with high probability. We formalize this problem and show assumptions under which it admits to efficient approximation algorithms. We verify these assumptions and validate the performance of our algorithms on real data.

[Approximate Heavily-Constrained Learning with Lagrange Multiplier Models](#)

- Harikrishna Narasimhan, Andrew Cotter, Yichen Zhou, Serena Wang, Wenshuo Guo
- abstract@[open-review](#): In machine learning applications such as ranking fairness or fairness over intersectional groups, one often encounters optimization problems with an extremely large number of constraints. In particular, with ranking fairness tasks, there may even be a variable number of constraints, e.g. one for each query in the training set. In these cases, the standard approach of optimizing a Lagrangian while maintaining one Lagrange multiplier per constraint may no longer be practical. Our proposal is to associate a feature vector with each constraint, and to learn a ``multiplier model'' that maps each such vector to the corresponding Lagrange multiplier. We prove optimality, approximate feasibility and generalization guarantees under assumptions on the flexibility of the multiplier model, and empirically demonstrate that our method is effective on real-world case studies.

[Faster Randomized Infeasible Interior Point Methods for Tall/Wide Linear Programs](#)

- Agniva Chowdhury, Palma London, Haim Avron, Petros Drineas
- abstract@[open-review](#): Linear programming (LP) is used in many machine learning applications, such as $\|\cdot\|_1$ -regularized SVMs, basis pursuit, nonnegative matrix factorization, etc. Interior Point Methods (IPMs) are one of the most popular methods to solve LPs both in theory and in practice. Their underlying complexity is dominated by the cost of solving a system of linear equations at each iteration. In this paper, we consider \{infeasible\} IPMs for the special case where the number of variables is much larger than the number of constraints (i.e., wide), or vice-versa (i.e., tall) by taking the dual. Using tools from Randomized Linear Algebra, we present a preconditioning technique that, when combined with the Conjugate Gradient iterative solver, provably guarantees that infeasible IPM algorithms (suitably modified to account for the error incurred by the approximate solver), converge to a feasible, approximately optimal solution, without increasing their iteration complexity. Our empirical evaluations verify our theoretical results on both real and synthetic data.

[Sliding Window Algorithms for k-Clustering Problems](#)

- Michele Borassi, Alessandro Epasto, Silvio Lattanzi, Sergei Vassilvitskii, Morteza Zadimoghaddam
- abstract@[open-review](#): The sliding window model of computation captures scenarios in which data is arriving continuously, but only the latest $\$w\$$ elements should be used for analysis. The goal is to design algorithms that update the solution efficiently with each arrival rather than recomputing it from scratch. In this work, we focus on $\$k\$$ -clustering problems such as $\$k\$$ -means and $\$k\$$ -median. In this setting, we provide simple and practical algorithms that offer stronger performance guarantees than previous results. Empirically, we show that our methods store only a small fraction of the data, are orders of magnitude faster, and find solutions with costs only slightly higher than those returned by algorithms with access to the full dataset.

[AdaShare: Learning What To Share For Efficient Deep Multi-Task Learning](#)

- Ximeng Sun, Rameswar Panda, Rogerio Feris, Kate Saenko

- abstract@[open-review](#): Multi-task learning is an open and challenging problem in computer vision. The typical way of conducting multi-task learning with deep neural networks is either through handcrafted schemes that share all initial layers and branch out at an adhoc point, or through separate task-specific networks with an additional feature sharing/fusion mechanism. Unlike existing methods, we propose an adaptive sharing approach, called AdaShare, that decides what to share across which tasks to achieve the best recognition accuracy, while taking resource efficiency into account. Specifically, our main idea is to learn the sharing pattern through a task-specific policy that selectively chooses which layers to execute for a given task in the multi-task network. We efficiently optimize the task-specific policy jointly with the network weights, using standard back-propagation. Experiments on several challenging and diverse benchmark datasets with a variable number of tasks well demonstrate the efficacy of our approach over state-of-the-art methods. Project page: <https://cs-people.bu.edu/sunxm/AdaShare/project.html>

[Approximate Cross-Validation for Structured Models](#)

- Soumya Ghosh, Will Stephenson, Tin D. Nguyen, Sameer Deshpande, Tamara Broderick
- abstract@[open-review](#): Many modern data analyses benefit from explicitly modeling dependence structure in data -- such as measurements across time or space, ordered words in a sentence, or genes in a genome. A gold standard evaluation technique is structured cross-validation (CV), which leaves out some data subset (such as data within a time interval or data in a geographic region) in each fold. But CV here can be prohibitively slow due to the need to re-run already-expensive learning algorithms many times. Previous work has shown approximate cross-validation (ACV) methods provide a fast and provably accurate alternative in the setting of empirical risk minimization. But this existing ACV work is restricted to simpler models by the assumptions that (i) data across CV folds are independent and (ii) an exact initial model fit is available. In structured data analyses, both these assumptions are often untrue. In the present work, we address (i) by extending ACV to CV schemes with dependence structure between the folds. To address (ii), we verify -- both theoretically and empirically -- that ACV quality deteriorates smoothly with noise in the initial fit. We demonstrate the accuracy and computational benefits of our proposed methods on a diverse set of real-world applications.

[Exemplar VAE: Linking Generative Models, Nearest Neighbor Retrieval, and Data Augmentation](#)

- Sajad Norouzi, David J. Fleet, Mohammad Norouzi
- abstract@[open-review](#): We introduce Exemplar VAEs, a family of generative models that bridge the gap between parametric and non-parametric, exemplar based generative models. Exemplar VAE is a variant of VAE with a non-parametric latent prior based on a Parzen window estimator. To sample from it, one first draws a random exemplar from a training set, then stochastically transforms that exemplar into a latent code and a new observation. We propose retrieval augmented training (RAT) as a way to speed up Exemplar VAE training by using approximate nearest neighbor search in the latent space to define a lower bound on log marginal likelihood. To enhance generalization, model parameters are learned using exemplar leave-one-out and subsampling. Experiments demonstrate the effectiveness of Exemplar VAEs on density estimation and representation learning. Importantly, generative data augmentation using Exemplar VAEs on permutation invariant MNIST and Fashion MNIST reduces classification error from 1.17% to 0.69% and from 8.56% to 8.16%.

[Debiased Contrastive Learning](#)

- Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, Stefanie Jegelka
- abstract@[open-review](#): A prominent technique for self-supervised representation learning has been to contrast semantically similar and dissimilar pairs of samples. Without access to labels, dissimilar (negative) points are typically taken to be randomly sampled datapoints, implicitly accepting that these points may, in reality, actually have the same label. Perhaps unsurprisingly, we observe that sampling negative examples from truly different labels improves performance, in a synthetic setting where labels are available. Motivated by this observation, we develop a debiased contrastive objective that corrects for the sampling of same-label datapoints, even without knowledge of the true labels. Empirically, the proposed objective consistently outperforms the state-of-the-art for representation learning in vision, language, and reinforcement learning benchmarks. Theoretically, we establish generalization bounds for the downstream classification task.

[UCSG-NET- Unsupervised Discovering of Constructive Solid Geometry Tree](#)

- Kacper Kania, Maciej Zieba, Tomasz Kajdanowicz
- abstract@[open-review](#): Signed distance field (SDF) is a prominent implicit representation of 3D meshes. Methods that are based on such representation achieved state-of-the-art 3D shape reconstruction quality. However, these methods struggle to reconstruct non-convex shapes. One remedy is to incorporate a constructive solid geometry framework (CSG) that represents a shape as a decomposition into primitives. It allows to embody a 3D shape of high complexity and non-convexity with a simple tree representation of Boolean operations. Nevertheless, existing approaches are supervised and require the entire CSG parse tree that is given upfront during the training process. On the contrary, we propose a model that extracts a CSG parse tree without any supervision - UCSG-Net. Our model predicts parameters of primitives and binarizes their SDF representation through differentiable indicator function. It is achieved jointly with discovering the structure of a Boolean operators tree. The model selects dynamically which operator combination over primitives leads to the reconstruction of high fidelity. We evaluate our method on 2D and 3D autoencoding tasks. We show that the predicted parse tree representation is interpretable and can be used in CAD software.

[Generalized Boosting](#)

- Arun Suggala, Bingbin Liu, Pradeep Ravikumar
- abstract@[open-review](#): Boosting is a widely used learning technique in machine learning for solving classification problems. In boosting, one predicts the label of an example using an ensemble of weak classifiers. While boosting has shown tremendous success on many classification problems involving tabular data, it performs poorly on complex classification tasks involving low-level features such as image classification tasks. This drawback stems from the fact that boosting builds an additive model of weak classifiers, each of which has very little predictive power. Often, the resulting additive models are not powerful enough to approximate the complex decision boundaries of real-world classification problems. In this work, we present a general framework for boosting where, similar to traditional boosting, we aim to boost the performance of a weak learner and transform it into a strong learner. However, unlike traditional boosting, our framework allows for more complex forms of aggregation of weak learners. In this work, we specifically focus on one form of aggregation - \emph{function composition}. We show that many popular greedy algorithms for learning deep neural networks (DNNs) can be derived from our framework using function compositions for aggregation. Moreover, we identify the drawbacks of these greedy algorithms and propose new algorithms that fix these issues. Using thorough empirical evaluation, we show that our learning algorithms have superior performance over traditional additive boosting algorithms, as well as existing greedy learning techniques for DNNs. An important feature of our algorithms is that they come with strong theoretical guarantees.

[COT-GAN: Generating Sequential Data via Causal Optimal Transport](#)

- Tianlin Xu, Li Kevin Wenliang, Michael Munn, Beatrice Acciaio
- abstract@[open-review](#): We introduce COT-GAN, an adversarial algorithm to train implicit generative models optimized for producing sequential data. The loss function of this algorithm is formulated using ideas from Causal Optimal Transport (COT), which combines classic optimal transport methods with an additional temporal causality constraint. Remarkably, we find that this causality condition provides a natural framework to parameterize the cost function that is learned by the discriminator as a robust (worst-case) distance, and an ideal mechanism for learning time dependent data distributions. Following Genevay et al. (2018), we also include an entropic penalization term which allows for the use of the Sinkhorn algorithm when computing the

optimal transport cost. Our experiments show effectiveness and stability of COT-GAN when generating both low- and high-dimensional time-series data. The success of the algorithm also relies on a new, improved version of the Sinkhorn divergence which demonstrates less bias in learning.

[Impossibility Results for Grammar-Compressed Linear Algebra](#)

- Amir Abboud, Arturs Backurs, Karl Bringmann, Marvin Kännemann
- abstract@[open-review](#): To handle vast amounts of data, it is natural and popular to compress vectors and matrices. When we compress a vector from size N down to size $n \ll N$, it certainly makes it easier to store and transmit efficiently, but does it also make it easier to process? In this paper we consider lossless compression schemes, and ask if we can run our computations on the compressed data as efficiently as if the original data was that small. That is, if an operation has time complexity $T(\text{input-size})$, can we perform it on the compressed representation in time $T(n)$ rather than $T(N)$? We consider the most basic linear algebra operations: inner product, matrix-vector multiplication, and matrix multiplication. In particular, given two compressed vectors, can we compute their inner product in time $O(n)$? Or perhaps we must decompress first and then multiply, spending $\Omega(N)$ time? The answer depends on the compression scheme. While for simple ones such as Run-Length-Encoding (RLE) the inner product can be done in $O(n)$ time, we prove that this is impossible for compressions from a richer class: essentially n^2 or even larger runtimes are needed in the worst case (under complexity assumptions). This is the class of \{grammar-compressions\} containing most popular methods such as the Lempel-Ziv family. These schemes are more compressing than the simple RLE, but alas, we prove that performing computations on them is much harder.

[Understanding spiking networks through convex optimization](#)

- Allan Mancio, Sander Keemink, Christian K. Machens
- abstract@[open-review](#): Neurons mainly communicate through spikes, and much effort has been spent to understand how the dynamics of spiking neural networks (SNNs) relates to their connectivity. Meanwhile, most major advances in machine learning have been made with simpler, rate-based networks, with SNNs only recently showing competitive results, largely thanks to transferring insights from rate to spiking networks. However, it is still an open question exactly which computations SNNs perform. Recently, the time-averaged firing rates of several SNNs were shown to yield the solutions to convex optimization problems. Here we turn these findings around and show that virtually all inhibition-dominated SNNs can be understood through the lens of convex optimization, with network connectivity, timescales, and firing thresholds being intricately linked to the parameters of underlying convex optimization problems. This approach yields new, geometric insights into the computations performed by spiking networks. In particular, we establish a class of SNNs whose instantaneous output provides a solution to linear or quadratic programming problems, and we thereby reveal their input-output mapping. Using these insights, we derive local, supervised learning rules that can approximate given convex input-output functions, and we show that the resulting networks are consistent with many features from biological networks, such as low firing rates, irregular firing, E/I balance, and robustness to perturbations and synaptic delays.

[Better Full-Matrix Regret via Parameter-Free Online Learning](#)

- Ashok Cutkosky
- abstract@[open-review](#): We provide online convex optimization algorithms that guarantee improved full-matrix regret bounds. These algorithms extend prior work in several ways. First, we seamlessly allow for the incorporation of constraints without requiring unknown oracle-tuning for any learning rate parameters. Second, we improve the regret of the full-matrix AdaGrad algorithm by suggesting a better learning rate value and showing how to tune the learning rate to this value on-the-fly. Third, all our bounds are obtained via a general framework for constructing regret bounds that depend on an arbitrary sequence of norms.

[Large-Scale Methods for Distributionally Robust Optimization](#)

- Daniel Levy, Yair Carmon, John C. Duchi, Aaron Sidford
- abstract@[open-review](#): We propose and analyze algorithms for distributionally robust optimization of convex losses with conditional value at risk (CVaR) and χ^2 divergence uncertainty sets. We prove that our algorithms require a number of gradient evaluations independent of training set size and number of parameters, making them suitable for large-scale applications. For χ^2 uncertainty sets these are the first such guarantees in the literature, and for CVaR our guarantees scale linearly in the uncertainty level rather than quadratically as in previous work. We also provide lower bounds proving the worst-case optimality of our algorithms for CVaR and a penalized version of the χ^2 problem. Our primary technical contributions are novel bounds on the bias of batch robust risk estimation and the variance of a multilevel Monte Carlo gradient estimator due to [Blanchet & Glynn, 2015]. Experiments on MNIST and ImageNet confirm the theoretical scaling of our algorithms, which are 9-36 times more efficient than full-batch methods.

[Analysis and Design of Thompson Sampling for Stochastic Partial Monitoring](#)

- Taira Tsuchiya, Junya Honda, Masashi Sugiyama
- abstract@[open-review](#): We investigate finite stochastic partial monitoring, which is a general model for sequential learning with limited feedback. While Thompson sampling is one of the most promising algorithms on a variety of online decision-making problems, its properties for stochastic partial monitoring have not been theoretically investigated, and the existing algorithm relies on a heuristic approximation of the posterior distribution. To mitigate these problems, we present a novel Thompson-sampling-based algorithm, which enables us to exactly sample the target parameter from the posterior distribution. Besides, we prove that the new algorithm achieves the logarithmic problem-dependent expected pseudo-regret $\mathcal{O}(\log T)$ for a linearized variant of the problem with local observability. This result is the first regret bound of Thompson sampling for partial monitoring, which also becomes the first logarithmic regret bound of Thompson sampling for linear bandits.

[Bandit Linear Control](#)

- Asaf Cassel, Tomer Koren
- abstract@[open-review](#): We consider the problem of controlling a known linear dynamical system under stochastic noise, adversarially chosen costs, and bandit feedback. Unlike the full feedback setting where the entire cost function is revealed after each decision, here only the cost incurred by the learner is observed. We present a new and efficient algorithm that, for strongly convex and smooth costs, obtains regret that grows with the square root of the time horizon T . We also give extensions of this result to general convex, possibly non-smooth costs, and to non-stochastic system noise. A key component of our algorithm is a new technique for addressing bandit optimization of loss functions with memory.

[Refactoring Policy for Compositional Generalizability using Self-Supervised Object Proposals](#)

- Tongzhou Mu, Jiayuan Gu, Zhiwei Jia, Hao Tang, Hao Su
- abstract@[open-review](#): We study how to learn a policy with compositional generalizability. We propose a two-stage framework, which refactors a high-reward teacher policy into a generalizable student policy with strong inductive bias. Particularly, we implement an object-centric GNN-based student policy, whose input objects are learned from images through self-supervised learning. Empirically, we evaluate our approach on four difficult tasks that require compositional generalizability, and achieve superior performance compared to baselines.

[PEP: Parameter Ensembling by Perturbation](#)

- Alireza Mehrtash, Purang Abolmaesumi, Polina Golland, Tina Kapur, Demian Wassermann, William Wells
- abstract@[open-review](#): Ensembling is now recognized as an effective approach for increasing the predictive performance and calibration of deep networks. We introduce a new approach, Parameter Ensembling by Perturbation (PEP), that constructs an ensemble of parameter values as random perturbations of the optimal parameter set from training by a Gaussian with a single variance parameter. The variance is chosen to maximize the log-likelihood of the ensemble average ($\delta \cdot f$) on the validation data set. Empirically, and perhaps surprisingly, $\delta \cdot f$ has a well-defined maximum as the variance grows from zero (which corresponds to the baseline model). Conveniently, calibration level of predictions also tends to grow favorably until the peak of $\delta \cdot f$ is reached. In most experiments, PEP provides a small improvement in performance, and, in some cases, a substantial improvement in empirical calibration. We show that this "PEP effect" (the gain in log-likelihood) is related to the mean curvature of the likelihood function and the empirical Fisher information. Experiments on ImageNet pre-trained networks including ResNet, DenseNet, and Inception showed improved calibration and likelihood. We further observed a mild improvement in classification accuracy on these networks. Experiments on classification benchmarks such as MNIST and CIFAR-10 showed improved calibration and likelihood, as well as the relationship between the PEP effect and overfitting; this demonstrates that PEP can be used to probe the level of overfitting that occurred during training. In general, no special training procedure or network architecture is needed, and in the case of pre-trained networks, no additional training is needed.

[Theoretical Insights Into Multiclass Classification: A High-dimensional Asymptotic View](#)

- Christos Thrampoulidis, Samet Oymak, Mahdi Soltanolkotabi
- abstract@[open-review](#): Contemporary machine learning applications often involve classification tasks with many classes. Despite their extensive use, a precise understanding of the statistical properties and behavior of classification algorithms is still missing, especially in modern regimes where the number of classes is rather large. In this paper, we take a step in this direction by providing the first asymptotically precise analysis of linear multiclass classification. Our theoretical analysis allows us to precisely characterize how the test error varies over different training algorithms, data distributions, problem dimensions as well as number of classes, inter/intra class correlations and class priors. Specifically, our analysis reveals that the classification accuracy is highly distribution-dependent with different algorithms achieving optimal performance for different data distributions and/or training/features sizes. Unlike linear regression/binary classification, the test error in multiclass classification relies on intricate functions of the trained model (e.g., correlation between some of the trained weights) whose asymptotic behavior is difficult to characterize. This challenge is already present in simple classifiers, such as those minimizing a square loss. Our novel theoretical techniques allow us to overcome some of these challenges. The insights gained may pave the way for a precise understanding of other classification algorithms beyond those studied in this paper.

[Adversarial Example Games](#)

- Joey Bose, Gauthier Gidel, Hugo Berard, Andre Cianflone, Pascal Vincent, Simon Lacoste-Julien, Will Hamilton
- abstract@[open-review](#): The existence of adversarial examples capable of fooling trained neural network classifiers calls for a much better understanding of possible attacks to guide the development of safeguards against them. This includes attack methods in the challenging {\em non-interactive blackbox} setting, where adversarial attacks are generated without any access, including queries, to the target model. Prior attacks in this setting have relied mainly on algorithmic innovations derived from empirical observations (e.g., that momentum helps), lacking principled transferability guarantees. In this work, we provide a theoretical foundation for crafting transferable adversarial examples to entire hypothesis classes. We introduce \textit{Adversarial Example Games} (AEG), a framework that models the crafting of adversarial examples as a min-max game between a generator of attacks and a classifier. AEG provides a new way to design adversarial examples by adversarially training a generator and a classifier from a given hypothesis class (e.g., architecture). We prove that this game has an equilibrium, and that the optimal generator is able to craft adversarial examples that can attack any classifier from the corresponding hypothesis class. We demonstrate the efficacy of AEG on the MNIST and CIFAR-10 datasets, outperforming prior state-of-the-art approaches with an average relative improvement of \\$29.9\%\\$ and \\$47.2\%\\$ against undefended and robust models (Table \ref{table:q2} \& \ref{table:q3}) respectively.

[Residual Distillation: Towards Portable Deep Neural Networks without Shortcuts](#)

- Guilin Li, Junlei Zhang, Yunhe Wang, Chuanjian Liu, Matthias Tan, Yunfeng Lin, Wei Zhang, Jiashi Feng, Tong Zhang
- abstract@[open-review](#): By transferring both features and gradients between different layers, shortcut connections explored by ResNets allow us to effectively train very deep neural networks up to hundreds of layers. However, the additional computation costs induced by those shortcuts are often overlooked. For example, during online inference, the shortcuts in ResNet-50 account for about 40 percent of the entire memory usage on feature maps, because the features in the preceding layers cannot be released until the subsequent calculation is completed. In this work, for the first time, we consider training the CNN models with shortcuts and deploying them without. In particular, we propose a novel joint-training framework to train plain CNN by leveraging the gradients of the ResNet counterpart. During forward step, the feature maps of the early stages of plain CNN are passed through later stages of both itself and the ResNet counterpart to calculate the loss. During backpropagation, gradients calculated from a mixture of these two parts are used to update the plainCNN network to solve the gradient vanishing problem. Extensive experiments on ImageNet/CIFAR10/CIFAR100 demonstrate that the plainCNN network without shortcuts generated by our approach can achieve the same level of accuracy as that of the ResNet baseline while achieving about \$1.4\times\$ speed-up and \$1.25\times\$ memory reduction. We also verified the feature transferability of our ImageNet pretrained plain-CNN network by fine-tuning it on MIT 67 and Caltech 101. Our results show that the performance of the plain-CNN is slightly higher than that of its baseline ResNet-50 on these two datasets. The codes are in: \url{https://github.com/leoozy/JointRD_Neurips2020} \{https://github.com/leoozy/JointRD_Neurips2020\}

[Provably Efficient Neural Estimation of Structural Equation Models: An Adversarial Approach](#)

- Luofeng Liao, You-Lin Chen, Zhuoran Yang, Bo Dai, Mladen Kolar, Zhaoran Wang
- abstract@[open-review](#): Structural equation models (SEMs) are widely used in sciences, ranging from economics to psychology, to uncover causal relationships underlying a complex system under consideration and estimate structural parameters of interest. We study estimation in a class of generalized SEMs where the object of interest is defined as the solution to a linear operator equation. We formulate the linear operator equation as a min-max game, where both players are parameterized by neural networks (NNs), and learn the parameters of these neural networks using the stochastic gradient descent. We consider both 2-layer and multi-layer NNs with ReLU activation functions and prove global convergence in an overparametrized regime, where the number of neurons is diverging. The results are established using techniques from online learning and local linearization of NNs, and improve in several aspects the current state-of-the-art. For the first time we provide a tractable estimation procedure for SEMs based on NNs with provable convergence and without the need for sample splitting.

[Security Analysis of Safe and Seldonian Reinforcement Learning Algorithms](#)

- Pinar Ozisik, Philip S. Thomas
- abstract@[open-review](#): We analyze the extent to which existing methods rely on accurate training data for a specific class of reinforcement learning (RL) algorithms, known as Safe and Seldonian RL. We introduce a new measure of security to quantify the susceptibility to perturbations in training data by creating an attacker model that represents a worst-case analysis, and show that a couple of Seldonian RL methods are extremely sensitive to even a few data corruptions. We then introduce a new algorithm that is more robust against data corruptions, and demonstrate its usage in practice on some RL problems, including a grid-world and a diabetes treatment simulation.

[Learning to Play Sequential Games versus Unknown Opponents](#)

- Pier Giuseppe Sessa, Ilija Bogunovic, Maryam Kamgarpour, Andreas Krause
- abstract@[open-review](#): We consider a repeated sequential game between a learner, who plays first, and an opponent who responds to the chosen action. We seek to design strategies for the learner to successfully interact with the opponent. While most previous approaches consider known opponent models, we focus on the setting in which the opponent's model is unknown. To this end, we use kernel-based regularity assumptions to capture and exploit the structure in the opponent's response. We propose a novel algorithm for the learner when playing against an adversarial sequence of opponents. The algorithm combines ideas from bilevel optimization and online learning to effectively balance between exploration (learning about the opponent's model) and exploitation (selecting highly rewarding actions for the learner). Our results include algorithm's regret guarantees that depend on the regularity of the opponent's response and scale sublinearly with the number of game rounds. Moreover, we specialize our approach to repeated Stackelberg games, and empirically demonstrate its effectiveness in a traffic routing and wildlife conservation task.

[Further Analysis of Outlier Detection with Deep Generative Models](#)

- Ziyu Wang, Bin Dai, David Wipf, Jun Zhu
- abstract@[open-review](#): The recent, counter-intuitive discovery that deep generative models (DGMs) can frequently assign a higher likelihood to outliers has implications for both outlier detection applications as well as our overall understanding of generative modeling. In this work, we present a possible explanation for this phenomenon, starting from the observation that a model's typical set and high-density region may not coincide. From this vantage point we propose a novel outlier test, the empirical success of which suggests that the failure of existing likelihood-based outlier tests does not necessarily imply that the corresponding generative model is uncalibrated. We also conduct additional experiments to help disentangle the impact of low-level texture versus high-level semantics in differentiating outliers. In aggregate, these results suggest that modifications to the standard evaluation practices and benchmarks commonly applied in the literature are needed.

[Bridging Imagination and Reality for Model-Based Deep Reinforcement Learning](#)

- Guangxiang Zhu, Minghao Zhang, Honglak Lee, Chongjie Zhang
- abstract@[open-review](#): Sample efficiency has been one of the major challenges for deep reinforcement learning. Recently, model-based reinforcement learning has been proposed to address this challenge by performing planning on imaginary trajectories with a learned world model. However, world model learning may suffer from overfitting to training trajectories, and thus model-based value estimation and policy search will be prone to be sucked in an inferior local policy. In this paper, we propose a novel model-based reinforcement learning algorithm, called Bridging Reality and Dream (BIRD). It maximizes the mutual information between imaginary and real trajectories so that the policy improvement learned from imaginary trajectories can be easily generalized to real trajectories. We demonstrate that our approach improves sample efficiency of model-based planning, and achieves state-of-the-art performance on challenging visual control benchmarks.

[Neural Networks Learning and Memorization with \(almost\) no Over-Parameterization](#)

- Amit Daniely
- abstract@[open-review](#): Many results in recent years established polynomial time learnability of various models via neural networks algorithms (e.g. \cite{andoni2014learning, daniely2016toward, daniely2017sgd, cao2019generalization, ziwei2019polylogarithmic, zou2019improved, ma2019comparative, du2018gradient, arora2019fine, song2019quadratic, oymak2019towards, ge2019mildly, brutzkus2018sgd}). However, unless the model is linear separable \cite{brutzkus2018sgd}, or the activation is a polynomial \cite{ge2019mildly}, these results require very large networks -- much more than what is needed for the mere existence of a good predictor.

In this paper we prove that SGD on depth two neural networks can memorize samples, learn polynomials with bounded weights, and learn certain kernel spaces, with $\{\text{near optimal}\}$ network size, sample complexity, and runtime. In particular, we show that SGD on depth two network with $\tilde{O}(\frac{m}{d})$ hidden neurons (and hence $\tilde{O}(m)$ parameters) can memorize m random labeled points in \mathbb{S}^{d-1} .

[Exploiting Higher Order Smoothness in Derivative-free Optimization and Continuous Bandits](#)

- Arya Akhavan, Massimiliano Pontil, Alexandre Tsybakov
- abstract@[open-review](#): We address the problem of zero-order optimization of a strongly convex function. The goal is to find the minimizer of the function by a sequential exploration of its function values, under measurement noise. We study the impact of higher order smoothness properties of the function on the optimization error and on the online regret. To solve this problem we consider a randomized approximation of the projected gradient descent algorithm. The gradient is estimated by a randomized procedure involving two function evaluations and a smoothing kernel. We derive upper bounds for this algorithm both in the constrained and unconstrained settings and prove minimax lower bounds for any sequential search method. Our results imply that the zero-order algorithm is nearly optimal in terms of sample complexity and the problem parameters. Based on this algorithm, we also propose an estimator of the minimum value of the function achieving almost sharp oracle behavior. We compare our results with the state-of-the-art, highlighting a number of key improvements.

[Towards a Combinatorial Characterization of Bounded-Memory Learning](#)

- Alon Gonen, Shachar Lovett, Michal Moshkovitz
- abstract@[open-review](#): Combinatorial dimensions play an important role in the theory of machine learning. For example, VC dimension characterizes PAC learning, SQ dimension characterizes weak learning with statistical queries, and Littlestone dimension characterizes online learning. In this paper we aim to develop combinatorial dimensions that characterize bounded memory learning. We propose a candidate solution for the case of realizable strong learning under a known distribution, based on the SQ dimension of neighboring distributions. We prove both upper and lower bounds for our candidate solution, that match in some regime of parameters. This is the first characterization of strong learning under space constraints in any regime. In this parameter regime there is an equivalence between bounded memory and SQ learning. We conjecture that our characterization holds in a much wider regime of parameters.

[Chaos, Extremism and Optimism: Volume Analysis of Learning in Games](#)

- Yun Kuen Cheung, Georgios Piliouras
- abstract@[open-review](#): We perform volume analysis of Multiplicative Weights Updates (MWU) and its optimistic variant (OMWU) in zero-sum as well as coordination games. Our analysis provides new insights into these game/dynamical systems, which seem hard to achieve via the classical techniques within Computer Science and Machine Learning. First, we examine these dynamics not in their original space (simplex of actions) but in a dual space (aggregate payoffs of actions). Second, we explore how the volume of a set of initial conditions evolves over time when it is pushed forward according to the algorithm. This is reminiscent of approaches in evolutionary game theory where replicator dynamics, the continuous-time analogue of MWU, is known to preserve volume in all games. Interestingly, when we examine discrete-time dynamics, the choices of the game and the algorithm both play a critical role. So whereas MWU expands volume in zero-sum games and is thus Lyapunov chaotic, we show that OMWU contracts volume, providing an alternative understanding for its known convergent behavior. Yet, we also prove a no-free-lunch type of theorem, in the sense that when examining coordination games the roles are reversed. Using these tools, we prove two novel, rather negative properties of MWU in zero-sum games. (1) Extremism:

even in games with a unique fully-mixed Nash equilibrium, the system recurrently gets stuck near pure-strategy profiles, despite them being clearly unstable from game-theoretic perspective. (2) Unavoidability: given any set of good states (with a rather relaxed interpretation of “good” states), the system cannot avoid bad states indefinitely.

[On Regret with Multiple Best Arms](#)

- Yinglun Zhu, Robert Nowak
- abstract@[open-review](#): We study a regret minimization problem with the existence of multiple best/near-optimal arms in the multi-armed bandit setting. We consider the case when the number of arms/actions is comparable or much larger than the time horizon, and make no assumptions about the structure of the bandit instance. Our goal is to design algorithms that can automatically adapt to the unknown hardness of the problem, i.e., the number of best arms. Our setting captures many modern applications of bandit algorithms where the action space is enormous and the information about the underlying instance/structure is unavailable. We first propose an adaptive algorithm that is agnostic to the hardness level and theoretically derive its regret bound. We then prove a lower bound for our problem setting, which indicates: (1) no algorithm can be minimax optimal simultaneously over all hardness levels; and (2) our algorithm achieves a rate function that is Pareto optimal. With additional knowledge of the expected reward of the best arm, we propose another adaptive algorithm that is minimax optimal, up to polylog factors, over all hardness levels. Experimental results confirm our theoretical guarantees and show advantages of our algorithms over the previous state-of-the-art.

[Matrix Completion with Hierarchical Graph Side Information](#)

- Adel Elmahdy, Junhyung Ahn, Changho Suh, Soheil Mohajer
- abstract@[open-review](#): We consider a matrix completion problem that exploits social or item similarity graphs as side information. We develop a universal, parameter-free, and computationally efficient algorithm that starts with hierarchical graph clustering and then iteratively refines estimates both on graph clustering and matrix ratings. Under a hierarchical stochastic block model that well respects practically-relevant social graphs and a low-rank rating matrix model (to be detailed), we demonstrate that our algorithm achieves the information-theoretic limit on the number of observed matrix entries (i.e., optimal sample complexity) that is derived by maximum likelihood estimation together with a lower-bound impossibility result. One consequence of this result is that exploiting the hierarchical structure of social graphs yields a substantial gain in sample complexity relative to the one that simply identifies different groups without resorting to the relational structure across them. We conduct extensive experiments both on synthetic and real-world datasets to corroborate our theoretical results as well as to demonstrate significant performance improvements over other matrix completion algorithms that leverage graph side information.

[Is Long Horizon RL More Difficult Than Short Horizon RL?](#)

- Ruosong Wang, Simon S. Du, Lin Yang, Sham Kakade
- abstract@[open-review](#): Learning to plan for long horizons is a central challenge in episodic reinforcement learning problems. A fundamental question is to understand how the difficulty of the problem scales as the horizon increases. Here the natural measure of sample complexity is a normalized one: we are interested in the $\{\text{number of episodes}\}$ it takes to provably discover a policy whose value is $\$varepsilon$ near to that of the optimal value, where the value is measured by the $\{\text{normalized}\}$ cumulative reward in each episode. In a COLT 2018 open problem, Jiang and Agarwal conjectured that, for tabular, episodic reinforcement learning problems, there exists a sample complexity lower bound which exhibits a polynomial dependence on the horizon --- a conjecture which is consistent with all known sample complexity upper bounds. This work refutes this conjecture, proving that tabular, episodic reinforcement learning is possible with a sample complexity that scales only $\{\text{logarithmically}\}$ with the planning horizon. In other words, when the values are appropriately normalized (to lie in the unit interval), this results shows that long horizon RL is no more difficult than short horizon RL, at least in a minimax sense. Our analysis introduces two ideas: (i) the construction of an $\$varepsilon$ -net for near-optimal policies whose log-covering number scales only logarithmically with the planning horizon, and (ii) the Online Trajectory Synthesis algorithm, which adaptively evaluates all policies in a given policy class and enjoys a sample complexity that scales logarithmically with the cardinality of the given policy class. Both may be of independent interest.

[Hamiltonian Monte Carlo using an adjoint-differentiated Laplace approximation: Bayesian inference for latent Gaussian models and beyond](#)

- Charles Margossian, Aki Vehtari, Daniel Simpson, Raj Agrawal
- abstract@[open-review](#): Gaussian latent variable models are a key class of Bayesian hierarchical models with applications in many fields. Performing Bayesian inference on such models can be challenging as Markov chain Monte Carlo algorithms struggle with the geometry of the resulting posterior distribution and can be prohibitively slow. An alternative is to use a Laplace approximation to marginalize out the latent Gaussian variables and then integrate out the remaining hyperparameters using dynamic Hamiltonian Monte Carlo, a gradient-based Markov chain Monte Carlo sampler. To implement this scheme efficiently, we derive a novel adjoint method that propagates the minimal information needed to construct the gradient of the approximate marginal likelihood. This strategy yields a scalable differentiation method that is orders of magnitude faster than state of the art differentiation techniques when the hyperparameters are high dimensional. We prototype the method in the probabilistic programming framework Stan and test the utility of the embedded Laplace approximation on several models, including one where the dimension of the hyperparameter is $\hat{\wedge} 46,000$. Depending on the cases, the benefits can include an alleviation of the geometric pathologies that frustrate Hamiltonian Monte Carlo and a dramatic speed-up.

[Adversarial Learning for Robust Deep Clustering](#)

- Xu Yang, Cheng Deng, Kun Wei, Junchi Yan, Wei Liu
- abstract@[open-review](#): Deep clustering integrates embedding and clustering together to obtain the optimal nonlinear embedding space, which is more effective in real-world scenarios compared with conventional clustering methods. However, the robustness of the clustering network is prone to being attenuated especially when it encounters an adversarial attack. A small perturbation in the embedding space will lead to diverse clustering results since the labels are absent. In this paper, we propose a robust deep clustering method based on adversarial learning. Specifically, we first attempt to define adversarial samples in the embedding space for the clustering network. Meanwhile, we devise an adversarial attack strategy to explore samples that easily fool the clustering layers but do not impact the performance of the deep embedding. We then provide a simple yet efficient defense algorithm to improve the robustness of the clustering network. Experimental results on two popular datasets show that the proposed adversarial learning method can significantly enhance the robustness and further improve the overall clustering performance. Particularly, the proposed method is generally applicable to multiple existing clustering frameworks to boost their robustness. The source code is available at <https://github.com/xdxuyang/ALRDC>.

[Learning Mutational Semantics](#)

- Brian Hie, Ellen Zhong, Bryan Bryson, Bonnie Berger
- abstract@[open-review](#): In many natural domains, changing a small part of an entity can transform its semantics; for example, a single word change can alter the meaning of a sentence, or a single amino acid change can mutate a viral protein to escape antiviral treatment or immunity. Although identifying such mutations can be desirable (for example, therapeutic design that anticipates avenues of viral escape), the rules governing semantic change are often hard to quantify. Here, we introduce the problem of identifying mutations with a large effect on semantics, but where valid mutations are under complex

constraints (for example, English grammar or biological viability), which we refer to as constrained semantic change search (CSCS). We propose an unsupervised solution based on language models that simultaneously learn continuous latent representations. We report good empirical performance on CSCS of single-word mutations to news headlines, map a continuous semantic space of viral variation, and, notably, show unprecedented zero-shot prediction of single-residue escape mutations to key influenza and HIV proteins, suggesting a productive link between modeling natural language and pathogenic evolution.

[Learning to Learn Variational Semantic Memory](#)

- Xiantong Zhen, Yingjun Du, Huan Xiong, Qiang Qiu, Cees Snoek, Ling Shao
- abstract@[open-review](#): In this paper, we introduce variational semantic memory into meta-learning to acquire long-term knowledge for few-shot learning. The variational semantic memory accrues and stores semantic information for the probabilistic inference of class prototypes in a hierarchical Bayesian framework. The semantic memory is grown from scratch and gradually consolidated by absorbing information from tasks it experiences. By doing so, it is able to accumulate long-term, general knowledge that enables it to learn new concepts of objects. We formulate memory recall as the variational inference of a latent memory variable from addressed contents, which offers a principled way to adapt the knowledge to individual tasks. Our variational semantic memory, as a new long-term memory module, confers principled recall and update mechanisms that enable semantic information to be efficiently accrued and adapted for few-shot learning. Experiments demonstrate that the probabilistic modelling of prototypes achieves a more informative representation of object classes compared to deterministic vectors. The consistent new state-of-the-art performance on four benchmarks shows the benefit of variational semantic memory in boosting few-shot recognition.

[Myersonian Regression](#)

- Allen Liu, Renato Leme, Jon Schneider
- abstract@[open-review](#): Motivated by pricing applications in online advertising, we study a variant of linear regression with a discontinuous loss function that we term Myersonian regression. In this variant, we wish to find a linear function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ that well approximates a set of points $(x_i, v_i) \in \mathbb{R}^d \times [0, 1]$ in the following sense: we receive a loss of v_i when $f(x_i) > v_i$ and a loss of $|v_i - f(x_i)|$ when $f(x_i) \leq v_i$. This arises naturally in the economic application of designing a pricing policy for differentiated items (where the loss is the gap between the performance of our policy and the optimal Myerson prices).

We show that Myersonian regression is NP-hard to solve exactly and furthermore that no fully polynomial-time approximation scheme exists for Myersonian regression conditioned on the Exponential Time Hypothesis being true. In contrast to this, we demonstrate a polynomial-time approximation scheme for Myersonian regression that obtains an ϵ additive approximation to the optimal possible revenue and can be computed in time $O(\exp(\mathrm{poly}(1/\epsilon))\mathrm{poly}(m, n))$. We show that this algorithm is stable and generalizes well over distributions of samples.

[Learnability with Indirect Supervision Signals](#)

- Kaifu Wang, Qiang Ning, Dan Roth
- abstract@[open-review](#): Learning from indirect supervision signals is important in real-world AI applications when, often, gold labels are missing or too costly. In this paper, we develop a unified theoretical framework for multi-class classification when the supervision is provided by a variable that contains nonzero mutual information with the gold label. The nature of this problem is determined by (i) the transition probability from the gold labels to the indirect supervision variables and (ii) the learner's prior knowledge about the transition. Our framework relaxes assumptions made in the literature, and supports learning with unknown, non-invertible and instance-dependent transitions. Our theory introduces a novel concept called *separation*, which characterizes the learnability and generalization bounds. We also demonstrate the application of our framework via concrete novel results in a variety of learning scenarios such as learning with superset annotations and joint supervision signals.

[Towards Safe Policy Improvement for Non-Stationary MDPs](#)

- Yash Chandak, Scott Jordan, Georgios Theocharous, Martha White, Philip S. Thomas
- abstract@[open-review](#): Many real-world sequential decision-making problems involve critical systems with financial risks and human-life risks. While several works in the past have proposed methods that are safe for deployment, they assume that the underlying problem is stationary. However, many real-world problems of interest exhibit non-stationarity, and when stakes are high, the cost associated with a false stationarity assumption may be unacceptable. We take the first steps towards ensuring safety, with high confidence, for smoothly-varying non-stationary decision problems. Our proposed method extends a type of safe algorithm, called a Seldonian algorithm, through a synthesis of model-free reinforcement learning with time-series analysis. Safety is ensured using sequential hypothesis testing of a policy's forecasted performance, and confidence intervals are obtained using wild bootstrap.

[Finer Metagenomic Reconstruction via Biodiversity Optimization](#)

- Simon Foucart, David Koslicki
- abstract@[open-review](#): When analyzing communities of microorganisms from their sequenced DNA, an important task is taxonomic profiling: enumerating the presence and relative abundance of all organisms, or merely of all taxa, contained in the sample. This task can be tackled via compressive-sensing-based approaches, which favor communities featuring the fewest organisms among those consistent with the observed DNA data. Despite their successes, these parsimonious approaches sometimes conflict with biological realism by overlooking organism similarities. Here, we leverage a recently developed notion of biological diversity that simultaneously accounts for organism similarities and retains the optimization strategy underlying compressive-sensing-based approaches. We demonstrate that minimizing biological diversity still produces sparse taxonomic profiles and we experimentally validate superiority to existing compressive-sensing-based approaches. Despite showing that the objective function is almost never convex and often concave, generally yielding NP-hard problems, we exhibit ways of representing organism similarities for which minimizing diversity can be performed via a sequence of linear programs guaranteed to decrease diversity. Better yet, when biological similarity is quantified by k-mer co-occurrence (a popular notion in bioinformatics), minimizing diversity actually reduces to one linear program that can utilize multiple k-mer sizes to enhance performance. In proof-of-concept experiments, we verify that the latter procedure can lead to significant gains when taxonomically profiling a metagenomic sample, both in terms of reconstruction accuracy and computational performance.

[Causal Discovery in Physical Systems from Videos](#)

- Yunzhu Li, Antonio Torralba, Anima Anandkumar, Dieter Fox, Animesh Garg
- abstract@[open-review](#): Causal discovery is at the core of human cognition. It enables us to reason about the environment and make counterfactual predictions about unseen scenarios that can vastly differ from our previous experiences. We consider the task of causal discovery from videos in an end-to-end fashion without supervision on the ground-truth graph structure. In particular, our goal is to discover the structural dependencies among environmental and object variables: inferring the type and strength of interactions that have a causal effect on the behavior of the dynamical system. Our model consists of (a) a perception module that extracts a semantically meaningful and temporally consistent keypoint representation from images, (b) an inference module for determining the graph distribution induced by the detected keypoints, and (c) a dynamics module that can predict the future by conditioning on the inferred graph. We assume access to different configurations and environmental conditions, i.e., data from unknown interventions on the underlying system; thus, we can hope to discover the correct underlying causal graph without explicit interventions. We evaluate our method in a planar multi-body interaction environment and scenarios involving fabrics of different shapes like shirts and pants. Experiments demonstrate that our model can correctly

identify the interactions from a short sequence of images and make long-term future predictions. The causal structure assumed by the model also allows it to make counterfactual predictions and extrapolate to systems of unseen interaction graphs or graphs of various sizes.

[Glyph: Fast and Accurately Training Deep Neural Networks on Encrypted Data](#)

- Qian Lou, Bo Feng, Geoffrey Charles Fox, Lei Jiang
- abstract@[open-review](#): Because of the lack of expertise, to gain benefits from their data, average users have to upload their private data to cloud servers they may not trust. Due to legal or privacy constraints, most users are willing to contribute only their encrypted data, and lack interests or resources to join deep neural network (DNN) training in cloud. To train a DNN on encrypted data in a completely non-interactive way, a recent work proposes a fully homomorphic encryption (FHE)-based technique implementing all activations by \textit{Brakerski-Gentry-Vaikuntanathan} (BGV)-based lookup tables. However, such inefficient lookup-table-based activations significantly prolong private training latency of DNNs. In this paper, we propose, Glyph, an FHE-based technique to fast and accurately train DNNs on encrypted data by switching between TFHE (Fast Fully Homomorphic Encryption over the Torus) and BGV cryptosystems. Glyph uses logic-operation-friendly TFHE to implement nonlinear activations, while adopts vectorial-arithmetic-friendly BGV to perform multiply-accumulations (MACs). Glyph further applies transfer learning on DNN training to improve test accuracy and reduce the number of MACs between ciphertext and ciphertext in convolutional layers. Our experimental results show Glyph obtains state-of-the-art accuracy, and reduces training latency by 69%~99% over prior FHE-based privacy-preserving techniques on encrypted datasets.

[Smoothed Analysis of Online and Differentially Private Learning](#)

- Nika Haghtalab, Tim Roughgarden, Abhishek Shetty
- abstract@[open-review](#): Practical and pervasive needs for robustness and privacy in algorithms have inspired the design of online adversarial and differentially private learning algorithms. The primary quantity that characterizes learnability in these settings is the Littlestone dimension of the class of hypotheses [Ben-David et al., 2009, Alon et al., 2019]. This characterization is often interpreted as an impossibility result because classes such as linear thresholds and neural networks have infinite Littlestone dimension. In this paper, we apply the framework of smoothed analysis [Spielman and Teng, 2004], in which adversarially chosen inputs are perturbed slightly by nature. We show that fundamentally stronger regret and error guarantees are possible with smoothed adversaries than with worst-case adversaries. In particular, we obtain regret and privacy error bounds that depend only on the VC dimension and the bracketing number of a hypothesis class, and on the magnitudes of the perturbations.

[Self-Paced Deep Reinforcement Learning](#)

- Pascal Klink, Carlo D'Eramo, Jan R. Peters, Joni Pajarinen
- abstract@[open-review](#): Curriculum reinforcement learning (CRL) improves the learning speed and stability of an agent by exposing it to a tailored series of tasks throughout learning. Despite empirical successes, an open question in CRL is how to automatically generate a curriculum for a given reinforcement learning (RL) agent, avoiding manual design. In this paper, we propose an answer by interpreting the curriculum generation as an inference problem, where distributions over tasks are progressively learned to approach the target task. This approach leads to an automatic curriculum generation, whose pace is controlled by the agent, with solid theoretical motivation and easily integrated with deep RL algorithms. In the conducted experiments, the curricula generated with the proposed algorithm significantly improve learning performance across several environments and deep RL algorithms, matching or outperforming state-of-the-art existing CRL algorithms.

[Kalman Filtering Attention for User Behavior Modeling in CTR Prediction](#)

- Hu Liu, Jing LU, Xiwei Zhao, Sulong Xu, Hao Peng, Yutong Liu, Zehua Zhang, Jian Li, Junsheng Jin, Yongjun Bao, Weipeng Yan
- abstract@[open-review](#): Click-through rate (CTR) prediction is one of the fundamental tasks for e-commerce search engines. As search becomes more personalized, it is necessary to capture the user interest from rich behavior data. Existing user behavior modeling algorithms develop different attention mechanisms to emphasize query-relevant behaviors and suppress irrelevant ones. Despite being extensively studied, these attentions still suffer from two limitations. First, conventional attentions mostly limit the attention field only to a single user's behaviors, which is not suitable in e-commerce where users often hunt for new demands that are irrelevant to any historical behaviors. Second, these attentions are usually biased towards frequent behaviors, which is unreasonable since high frequency does not necessarily indicate great importance. To tackle the two limitations, we propose a novel attention mechanism, termed Kalman Filtering Attention (KFAtt), that considers the weighted pooling in attention as a maximum a posteriori (MAP) estimation. By incorporating a priori, KFAtt resorts to global statistics when few user behaviors are relevant. Moreover, a frequency capping mechanism is incorporated to correct the bias towards frequent behaviors. Offline experiments on both benchmark and a 10 billion scale real production dataset, together with an Online A/B test, show that KFAtt outperforms all compared state-of-the-arts. KFAtt has been deployed in the ranking system of JD.com, one of the largest B2C e-commerce websites in China, serving the main traffic of hundreds of millions of active users.

[Towards Maximizing the Representation Gap between In-Domain & Out-of-Distribution Examples](#)

- Jay Nandy, Wynne Hsu, Mong Li Lee
- abstract@[open-review](#): Among existing uncertainty estimation approaches, Dirichlet Prior Network (DPN) distinctly models different predictive uncertainty types. However, for in-domain examples with high data uncertainties among multiple classes, even a DPN model often produces indistinguishable representations from the out-of-distribution (OOD) examples, compromising their OOD detection performance. We address this shortcoming by proposing a novel loss function for DPN to maximize the representation gap between in-domain and OOD examples. Experimental results demonstrate that our proposed approach consistently improves OOD detection performance.

[Fully Convolutional Mesh Autoencoder using Efficient Spatially Varying Kernels](#)

- Yi Zhou, Chenglei Wu, Zimo Li, Chen Cao, Yuting Ye, Jason Saragih, Hao Li, Yaser Sheikh
- abstract@[open-review](#): Learning latent representations of registered meshes is useful for many 3D tasks. Techniques have recently shifted to neural mesh autoencoders. Although they demonstrate higher precision than traditional methods, they remain unable to capture fine-grained deformations. Furthermore, these methods can only be applied to a template-specific surface mesh, and is not applicable to more general meshes, like tetrahedrons and non-manifold meshes. While more general graph convolution methods can be employed, they lack performance in reconstruction precision and require higher memory usage. In this paper, we propose a non-template-specific fully convolutional mesh autoencoder for arbitrary registered mesh data. It is enabled by our novel convolution and (un)pooling operators learned with globally shared weights and locally varying coefficients which can efficiently capture the spatially varying contents presented by irregular mesh connections. Our model outperforms state-of-the-art methods on reconstruction accuracy. In addition, the latent codes of our network are fully localized thanks to the fully convolutional structure, and thus have much higher interpolation capability than many traditional 3D mesh generation models.

[GNNGuard: Defending Graph Neural Networks against Adversarial Attacks](#)

- Xiang Zhang, Marinka Zitnik
- abstract@[open-review](#): Deep learning methods for graphs achieve remarkable performance on many tasks. However, despite the proliferation of such methods and their success, recent findings indicate that small, unnoticeable perturbations of graph structure can catastrophically reduce performance of

even the strongest and most popular Graph Neural Networks (GNNs). Here, we develop GNNGuard, a general defense approach against a variety of training-time attacks that perturb the discrete graph structure. GNNGuard can be straightforwardly incorporated into any GNN. Its core principle is to detect and quantify the relationship between the graph structure and node features, if one exists, and then exploit that relationship to mitigate the negative effects of the attack. GNNGuard learns how to best assign higher weights to edges connecting similar nodes while pruning edges between unrelated nodes. The revised edges then allow the underlying GNN to robustly propagate neural messages in the graph. GNNGuard introduces two novel components, the neighbor importance estimation, and the layer-wise graph memory, and we show empirically that both components are necessary for a successful defense. Across five GNNs, three defense methods, and four datasets, including a challenging human disease graph, experiments show that GNNGuard outperforms existing defense approaches by 15.3% on average. Remarkably, GNNGuard can effectively restore state-of-the-art performance of GNNs in the face of various adversarial attacks, including targeted and non-targeted attacks, and can defend against attacks on heterophily graphs.

[Geo-PIFu: Geometry and Pixel Aligned Implicit Functions for Single-view Human Reconstruction](#)

- Tong He, John Collomosse, Hailin Jin, Stefano Soatto
- abstract@[open-review](#): We propose Geo-PIFu, a method to recover a 3D mesh from a monocular color image of a clothed person. Our method is based on a deep implicit function-based representation to learn latent voxel features using a structure-aware 3D U-Net, to constrain the model in two ways: first, to resolve feature ambiguities in query point encoding, second, to serve as a coarse human shape proxy to regularize the high-resolution mesh and encourage global shape regularity. We show that, by both encoding query points and constraining global shape using latent voxel features, the reconstruction we obtain for clothed human meshes exhibits less shape distortion and improved surface details compared to competing methods. We evaluate Geo-PIFu on a recent human mesh public dataset that is 10x larger than the private commercial dataset used in PIFu and previous derivative work. On average, we exceed the state of the art by 42.7% reduction in Chamfer and Point-to-Surface Distances, and 19.4% reduction in normal estimation errors.

[Optimal visual search based on a model of target detectability in natural images](#)

- Shima Rashidi, Krista Ehinger, Andrew Turpin, Lars Kulik
- abstract@[open-review](#): To analyse visual systems, the concept of an ideal observer promises an optimal response for a given task. Bayesian ideal observers can provide optimal responses under uncertainty, if they are given the true distributions as input. In visual search tasks, prior studies have used signal to noise ratio (SNR) or psychophysics experiments to set the distributional parameters for simple targets on backgrounds with known patterns, however these methods do not easily translate to complex targets on natural scenes. Here, we develop a model of target detectability in natural images to estimate the parameters of target-present and target-absent distributions for a visual search task. We present a novel approach for approximating the foveated detectability of a known target in natural backgrounds based on biological aspects of human visual system. Our model considers both the uncertainty about target position and the visual system's variability due to its reduced performance in the periphery compared to the fovea. Our automated prediction algorithm uses trained logistic regression as a post processing phase of a pre-trained deep neural network. Eye tracking data from 12 observers detecting targets on natural image backgrounds are used as ground truth to tune foveation parameters and evaluate the model, using cross-validation. Finally, the model of target detectability is used in a Bayesian ideal observer model of visual search, and compared to human search performance.

[Towards Convergence Rate Analysis of Random Forests for Classification](#)

- Wei Gao, Zhi-Hua Zhou
- abstract@[open-review](#): Random forests have been one of the successful ensemble algorithms in machine learning. The basic idea is to construct a large number of random trees individually and make prediction based on an average of their predictions. The great successes have attracted much attention on the consistency of random forests, mostly focusing on regression. This work takes one step towards convergence rates of random forests for classification. We present the first finite-sample rate $O(n^{-1/(8d+2)})$ on the convergence of pure random forests for classification, which can be improved to be of $O(n^{-1/(3.87d+2)})$ by considering the midpoint splitting mechanism. We introduce another variant of random forests, which follow Breiman's original random forests but with different mechanisms on splitting dimensions and positions. We get a convergence rate $O(n^{-\{1\}/(d+2)}(\ln n)^{\{1\}/(d+2)})$ for the variant of random forests, which reaches the minimax rate, except for a factor $(\ln n)^{\{1\}/(d+2)}$, of the optimal plug-in classifier under the L-Lipschitz assumption. We achieve tighter convergence rate $O(\sqrt{\ln n})$ under proper assumptions over structural data.

[List-Decodable Mean Estimation via Iterative Multi-Filtering](#)

- Ilias Diakonikolas, Daniel Kane, Daniel Kongsgaard
- abstract@[open-review](#): We study the problem of {\em list-decodable mean estimation} for bounded covariance distributions. Specifically, we are given a set \mathcal{T} of points in \mathbb{R}^d with the promise that an unknown α -fraction of points in \mathcal{T} , where $0 < \alpha < 1/2$, are drawn from an unknown mean and bounded covariance distribution D , and no assumptions are made on the remaining points. The goal is to output a small list of hypothesis vectors such that at least one of them is close to the mean of D .

We give the first practically viable estimator for this problem. In more detail, our algorithm is sample and computationally efficient, and achieves information-theoretically near-optimal error. While the only prior algorithm for this setting inherently relied on the ellipsoid method, our algorithm is iterative and only uses spectral techniques. Our main technical innovation is the design of a soft outlier removal procedure for high-dimensional heavy-tailed datasets with a majority of outliers.

[Exact Recovery of Mangled Clusters with Same-Cluster Queries](#)

- Marco Bressan, Nicolò Cesa-Bianchi, Silvio Lattanzi, Andrea Paudice
- abstract@[open-review](#): We study the cluster recovery problem in the semi-supervised active clustering framework. Given a finite set of input points, and an oracle revealing whether any two points lie in the same cluster, our goal is to recover all clusters exactly using as few queries as possible. To this end, we relax the spherical k -means cluster assumption of Ashtiani et al.\ to allow for arbitrary ellipsoidal clusters with margin. This removes the assumption that the clustering is center-based (i.e., defined through an optimization problem), and includes all those cases where spherical clusters are individually transformed by any combination of rotations, axis scalings, and point deletions. We show that, even in this much more general setting, it is still possible to recover the latent clustering exactly using a number of queries that scales only logarithmically with the number of input points. More precisely, we design an algorithm that, given n points to be partitioned into k clusters, uses $O(k^3 \ln k \ln n)$ oracle queries and $\tilde{O}(kn + k^3)$ time to recover the clustering with zero misclassification error. The $O(\cdot)$ notation hides an exponential dependence on the dimensionality of the clusters, which we show to be necessary thus characterizing the query complexity of the problem. Our algorithm is simple, easy to implement, and can also learn the clusters using low-stretch separators, a class of ellipsoids with additional theoretical guarantees. Experiments on large synthetic datasets confirm that we can reconstruct clusterings exactly and efficiently.

[Steady State Analysis of Episodic Reinforcement Learning](#)

- Huang Bojun
- abstract@[open-review](#): Reinforcement Learning (RL) tasks generally divide into two kinds: continual learning and episodic learning. The concept of steady state has played a foundational role in the continual setting, where unique steady-state distribution is typically presumed to exist in the task being studied, which enables principled conceptual framework as well as efficient data collection method for continual RL algorithms. On the other hand, the concept of steady state has been widely considered irrelevant for episodic RL tasks, in which the decision process terminates in finite time. Alternative

concepts, such as episode-wise visitation frequency, are used in episodic RL algorithms, which are not only inconsistent with their counterparts in continual RL, and also make it harder to design and analyze RL algorithms in the episodic setting. In this paper we proved that unique steady-state distributions pervasively exist in the learning environment of episodic learning tasks, and that the marginal distributions of the system state indeed approach to the steady state in essentially all episodic tasks. This observation supports an interestingly reversed mindset against conventional wisdom: While steady states are traditionally presumed to exist in continual learning and considered less relevant in episodic learning, it turns out they are guaranteed to exist for the latter under any behavior policy. We further developed interesting connections for important concepts that have been separately treated in episodic and continual RL. At the practical side, the existence of unique and approachable steady state implies a general, reliable, and efficient way to collect data in episodic RL algorithms. We applied this method to policy gradient algorithms, based on a new steady-state policy gradient theorem. We also proposed and experimentally evaluated a perturbation method to enforce faster convergence to steady state in real-world episodic RL tasks.

[Direct Feedback Alignment Scales to Modern Deep Learning Tasks and Architectures](#)

- Julien Launay, Iacopo Poli, FranÃ§ois Boniface, Florent Krzakala
- abstract@[open-review](#): Despite being the workhorse of deep learning, the backpropagation algorithm is no panacea. It enforces sequential layer updates, thus preventing efficient parallelization of the training process. Furthermore, its biological plausibility is being challenged. Alternative schemes have been devised; yet, under the constraint of synaptic asymmetry, none have scaled to modern deep learning tasks and architectures. Here, we challenge this perspective, and study the applicability of Direct Feedback Alignment (DFA) to neural view synthesis, recommender systems, geometric learning, and natural language processing. In contrast with previous studies limited to computer vision tasks, our findings show that it successfully trains a large range of state-of-the-art deep learning architectures, with performance close to fine-tuned backpropagation. When a larger gap between DFA and backpropagation exists, like in Transformers, we attribute this to a need to rethink common practices for large and complex architectures. At variance with common beliefs, our work supports that challenging tasks can be tackled in the absence of weight transport.

[Bayesian Optimization for Iterative Learning](#)

- Vu Nguyen, Sebastian Schulze, Michael Osborne
- abstract@[open-review](#): The performance of deep (reinforcement) learning systems crucially depends on the choice of hyperparameters. Their tuning is notoriously expensive, typically requiring an iterative training process to run for numerous steps to convergence. Traditional tuning algorithms only consider the final performance of hyperparameters acquired after many expensive iterations and ignore intermediate information from earlier training steps. In this paper, we present a Bayesian optimization(BO) approach which exploits the iterative structure of learning algorithms for efficient hyperparameter tuning. We propose to learn an evaluation function compressing learning progress at any stage of the training process into a single numeric score according to both training success and stability. Our BO framework is then trade-off the benefit of assessing a hyperparameter setting over additional training steps against their computation cost. We further increase model efficiency by selectively including scores from different training steps for any evaluated hyperparameter set. We demonstrate the efficiency of our algorithm by tuning hyperparameters for the training of deep reinforcement learning agents and convolutional neural networks. Our algorithm outperforms all existing baselines in identifying optimal hyperparameters in minimal time.

[Minimax Bounds for Generalized Linear Models](#)

- Kuan-Yun Lee, Thomas Courtade
- abstract@[open-review](#): We establish a new class of minimax prediction error bounds for generalized linear models. Our bounds significantly improve previous results when the design matrix is poorly structured, including natural cases where the matrix is wide or does not have full column rank. Apart from the typical L_2 risks, we study a class of entropic risks which recovers the usual L_2 prediction and estimation risks, and demonstrate that a tight analysis of Fisher information can uncover underlying structural dependency in terms of the spectrum of the design matrix. The minimax approach we take differs from the traditional metric entropy approach, and can be applied to many other settings.

[Projection Robust Wasserstein Distance and Riemannian Optimization](#)

- Tianyi Lin, Chenyou Fan, Nhat Ho, Marco Cuturi, Michael Jordan
- abstract@[open-review](#): Projection robust Wasserstein (PRW) distance, or Wasserstein projection pursuit (WPP), is a robust variant of the Wasserstein distance. Recent work suggests that this quantity is more robust than the standard Wasserstein distance, in particular when comparing probability measures in high-dimensions. However, it is ruled out for practical application because the optimization model is essentially non-convex and non-smooth which makes the computation intractable. Our contribution in this paper is to revisit the original motivation behind WPP/PRW, but take the hard route of showing that, despite its non-convexity and lack of nonsmoothness, and even despite some hardness results proved by \textit{Niles-2019-Estimation} in a minimax sense, the original formulation for PRW/WPP can be efficiently computed in practice using Riemannian optimization, yielding in relevant cases better behavior than its convex relaxation. More specifically, we provide three simple algorithms with solid theoretical guarantee on their complexity bound (one in the appendix), and demonstrate their effectiveness and efficiency by conducting extensive experiments on synthetic and real data. This paper provides a first step into a computational theory of the PRW distance and provides the links between optimal transport and Riemannian optimization.

[CoinDICE: Off-Policy Confidence Interval Estimation](#)

- Bo Dai, Ofir Nachum, Yinlam Chow, Lihong Li, Csaba Szepesvari, Dale Schuurmans
- abstract@[open-review](#): We study high-confidence behavior-agnostic off-policy evaluation in reinforcement learning, where the goal is to estimate a confidence interval on a target policy's value, given only access to a static experience dataset collected by unknown behavior policies. Starting from a function space embedding of the linear program formulation of the Q-function, we obtain an optimization problem with generalized estimating equation constraints. By applying the generalized empirical likelihood method to the resulting Lagrangian, we propose CoinDICE, a novel and efficient algorithm for computing confidence intervals. Theoretically, we prove the obtained confidence intervals are valid, in both asymptotic and finite-sample regimes. Empirically, we show in a variety of benchmarks that the confidence interval estimates are tighter and more accurate than existing methods.

[Simple and Fast Algorithm for Binary Integer and Online Linear Programming](#)

- Xiaocheng Li, Chunlin Sun, Yinyu Ye
- abstract@[open-review](#): In this paper, we develop a simple and fast online algorithm for solving a class of binary integer linear programs (LPs) arisen in the general resource allocation problem. The algorithm requires only one single pass through the input data and is free of doing any matrix inversion. It can be viewed as both an approximate algorithm for solving binary integer LPs and a fast algorithm for solving online LP problems. The algorithm is inspired by an equivalent form of the dual problem of the relaxed LP and it essentially performs (one-pass) projected stochastic subgradient descent in the dual space. We analyze the algorithm under two different models, stochastic input and random permutation, with minimal technical assumptions on the input data. The algorithm achieves $O(\sqrt{mn})$ expected regret under the stochastic input model and $O((m+\log n)\sqrt{n})$ expected regret under the random permutation model, and it achieves $O(m\sqrt{n})$ expected constraint violation under both models, where n is the number of decision variables and m is the number of constraints. In addition, we employ the notion of permutational Rademacher complexity and derive regret bounds for two earlier online LP algorithms for comparison. Both algorithms improve the regret bound with a factor of \sqrt{m} by paying

more computational cost. Furthermore, we demonstrate how to convert the possibly infeasible solution to a feasible one through a randomized procedure. Numerical experiments illustrate the general applicability and effectiveness of the algorithms.

[Learning Diverse and Discriminative Representations via the Principle of Maximal Coding Rate Reduction](#)

- Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, Yi Ma
- abstract@[open-review](#): To learn intrinsic low-dimensional structures from high-dimensional data that most discriminate between classes, we propose the principle of {\em Maximal Coding Rate Reduction} (\$\text{MCR}^2\$), an information-theoretic measure that maximizes the coding rate difference between the whole dataset and the sum of each individual class. We clarify its relationships with most existing frameworks such as cross-entropy, information bottleneck, information gain, contractive and contrastive learning, and provide theoretical guarantees for learning diverse and discriminative features. The coding rate can be accurately computed from finite samples of degenerate subspace-like distributions and can learn intrinsic representations in supervised, self-supervised, and unsupervised settings in a unified manner. Empirically, the representations learned using this principle alone are significantly more robust to label corruptions in classification than those using cross-entropy, and can lead to state-of-the-art results in clustering mixed data from self-learned invariant features.

[Learning Rich Rankings](#)

- Arjun Seshadri, Stephen Ragaini, Johan Ugander
- abstract@[open-review](#): Although the foundations of ranking are well established, the ranking literature has primarily been focused on simple, unimodal models, e.g. the Mallows and Plackett-Luce models, that define distributions centered around a single total ordering. Explicit mixture models have provided some tools for modelling multimodal ranking data, though learning such models from data is often difficult. In this work, we contribute a contextual repeated selection (CRS) model that leverages recent advances in choice modeling to bring a natural multimodality and richness to the rankings space. We provide rigorous theoretical guarantees for maximum likelihood estimation under the model through structure-dependent tail risk and expected risk bounds. As a by-product, we also furnish the first tight bounds on the expected risk of maximum likelihood estimators for the multinomial logit (MNL) choice model and the Plackett-Luce (PL) ranking model, as well as the first tail risk bound on the PL ranking model. The CRS model significantly outperforms existing methods for modeling real world ranking data in a variety of settings, from racing to rank choice voting.

[Color Visual Illusions: A Statistics-based Computational Model](#)

- Elad Hirsch, Ayellet Tal
- abstract@[open-review](#): Visual illusions may be explained by the likelihood of patches in real-world images, as argued by input-driven paradigms in Neuro-Science. However, neither the data nor the tools existed in the past to extensively support these explanations. The era of big data opens a new opportunity to study input-driven approaches. We introduce a tool that computes the likelihood of patches, given a large dataset to learn from. Given this tool, we present a model that supports the approach and explains lightness and color visual illusions in a unified manner. Furthermore, our model generates visual illusions in natural images, by applying the same tool, reversely.

[Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](#)

- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kästner, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, Douwe Kiela
- abstract@[open-review](#): Large pre-trained language models have been shown to store factual knowledge in their parameters, and achieve state-of-the-art results when fine-tuned on downstream NLP tasks. However, their ability to access and precisely manipulate knowledge is still limited, and hence on knowledge-intensive tasks, their performance lags behind task-specific architectures. Additionally, providing provenance for their decisions and updating their world knowledge remain open research problems. Pre-trained models with a differentiable access mechanism to explicit non-parametric memory can overcome this issue, but have so far been only investigated for extractive downstream tasks. We explore a general-purpose fine-tuning recipe for retrieval-augmented generation (RAG) -- models which combine pre-trained parametric and non-parametric memory for language generation. We introduce RAG models where the parametric memory is a pre-trained seq2seq model and the non-parametric memory is a dense vector index of Wikipedia, accessed with a pre-trained neural retriever. We compare two RAG formulations, one which conditions on the same retrieved passages across the whole generated sequence, the other can use different passages per token. We fine-tune and evaluate our models on a wide range of knowledge-intensive NLP tasks and set the state-of-the-art on three open domain QA tasks, outperforming parametric seq2seq models and task-specific retrieve-and-extract architectures. For language generation tasks, we find that RAG models generate more specific, diverse and factual language than a state-of-the-art parametric-only seq2seq baseline.

[Universal guarantees for decision tree induction via a higher-order splitting criterion](#)

- Guy Blanc, Neha Gupta, Jane Lange, Li-Yang Tan
- abstract@[open-review](#): We propose a simple extension of {\em top-down decision tree learning heuristics} such as ID3, C4.5, and CART. Our algorithm achieves provable guarantees for all target functions $f : \{-1,1\}^n \rightarrow \{-1,1\}$ with respect to the uniform distribution, circumventing impossibility results showing that existing heuristics fare poorly even for simple target functions. The crux of our extension is a new splitting criterion that takes into account the correlations between f and {\em small subsets} of its attributes. The splitting criteria of existing heuristics (e.g. Gini impurity and information gain), in contrast, are based solely on the correlations between f and its {\em individual} attributes.

Our algorithm satisfies the following guarantee: for all target functions $f : \{-1,1\}^n \rightarrow \{-1,1\}$, sizes $s \in \mathbb{N}$, and error parameters ϵ , it constructs a decision tree of size $s^{\tilde{O}(\log s)^2/\epsilon^2}$ that achieves error $O(\epsilon \log s) + \epsilon$, where $\log s$ denotes the error of the optimal size- s decision tree for f . A key technical notion that drives our analysis is the {\em noise stability} of f , a well-studied smoothness measure of f .

[Trade-offs and Guarantees of Adversarial Representation Learning for Information Obfuscation](#)

- Han Zhao, Jianfeng Chi, Yuan Tian, Geoffrey J. Gordon
- abstract@[open-review](#): Crowdsourced data used in machine learning services might carry sensitive information about attributes that users do not want to share. Various methods have been proposed to minimize the potential information leakage of sensitive attributes while maximizing the task accuracy. However, little is known about the theory behind these methods. In light of this gap, we develop a novel theoretical framework for attribute obfuscation. Under our framework, we propose a minimax optimization formulation to protect the given attribute and analyze its inference guarantees against worst-case adversaries. Meanwhile, there is a tension between minimizing information leakage and maximizing task accuracy. To understand this, we prove an information-theoretic lower bound to precisely characterize the fundamental trade-off between accuracy and information leakage. We conduct experiments on two real-world datasets to corroborate the inference guarantees and validate the inherent trade-offs therein. Our results indicate that, among several alternatives, the adversarial learning approach achieves the best trade-off in terms of attribute obfuscation and accuracy maximization.

[A Boolean Task Algebra for Reinforcement Learning](#)

- Geraud Nangue Tasse, Steven James, Benjamin Rosman

- abstract@[open-review](#): The ability to compose learned skills to solve new tasks is an important property for lifelong-learning agents. In this work we formalise the logical composition of tasks as a Boolean algebra. This allows us to formulate new tasks in terms of the negation, disjunction and conjunction of a set of base tasks. We then show that by learning goal-oriented value functions and restricting the transition dynamics of the tasks, an agent can solve these new tasks with no further learning. We prove that by composing these value functions in specific ways, we immediately recover the optimal policies for all tasks expressible under the Boolean algebra. We verify our approach in two domains---including a high-dimensional video game environment requiring function approximation---where an agent first learns a set of base skills, and then composes them to solve a super-exponential number of new tasks.

[Learning with Differentiable Perturbed Optimizers](#)

- Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, Francis Bach
- abstract@[open-review](#): Machine learning pipelines often rely on optimizers procedures to make discrete decisions (e.g., sorting, picking closest neighbors, or shortest paths). Although these discrete decisions are easily computed in a forward manner, they break the back-propagation of computational graphs. In order to expand the scope of learning problems that can be solved in an end-to-end fashion, we propose a systematic method to transform optimizers into operations that are differentiable and never locally constant. Our approach relies on stochastically perturbed optimizers, and can be used readily within existing solvers. Their derivatives can be evaluated efficiently, and smoothness tuned via the chosen noise amplitude. We also show how this framework can be connected to a family of losses developed in structured prediction, and give theoretical guarantees for their use in learning tasks. We demonstrate experimentally the performance of our approach on various tasks.

[Optimal Learning from Verified Training Data](#)

- Nicholas Bishop, Long Tran-Thanh, Enrico Gerding
- abstract@[open-review](#): Standard machine learning algorithms typically assume that data is sampled independently from the distribution of interest. In attempts to relax this assumption, fields such as adversarial learning typically assume that data is provided by an adversary, whose sole objective is to fool a learning algorithm. However, in reality, it is often the case that data comes from self-interested agents, with less malicious goals and intentions which lie somewhere between the two settings described above. To tackle this problem, we present a Stackelberg competition model for least squares regression, in which data is provided by agents who wish to achieve specific predictions for their data. Although the resulting optimisation problem is nonconvex, we derive an algorithm which converges globally, outperforming current approaches which only guarantee convergence to local optima. We also provide empirical results on two real-world datasets, the medical personal costs dataset and the red wine dataset, showcasing the performance of our algorithm relative to algorithms which are optimal under adversarial assumptions, outperforming the state of the art.

[Online Linear Optimization with Many Hints](#)

- Aditya Bhaskara, Ashok Cutkosky, Ravi Kumar, Manish Purohit
- abstract@[open-review](#): We study an online linear optimization (OLO) problem in which the learner is provided access to $\$K\$$ ``hint'' vectors in each round prior to making a decision. In this setting, we devise an algorithm that obtains logarithmic regret whenever there exists a convex combination of the $\$K\$$ hints that has positive correlation with the cost vectors. This significantly extends prior work that considered only the case $\$K=1\$$. To accomplish this, we develop a way to combine many arbitrary OLO algorithms to obtain regret only a logarithmically worse factor than the minimum regret of the original algorithms in hindsight; this result is of independent interest.

[Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification](#)

- Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, Lenka Zdeborová
- abstract@[open-review](#): We analyze in a closed form the learning dynamics of stochastic gradient descent (SGD) for a single layer neural network classifying a high-dimensional Gaussian mixture where each cluster is assigned one of two labels. This problem provides a prototype of a non-convex loss landscape with interpolating regimes and a large generalization gap. We define a particular stochastic process for which SGD can be extended to a continuous-time limit that we call stochastic gradient flow. In the full-batch limit we recover the standard gradient flow. We apply dynamical mean-field theory from statistical physics to track the dynamics of the algorithm in the high-dimensional limit via a self-consistent stochastic process. We explore the performance of the algorithm as a function of control parameters shedding light on how it navigates the loss landscape.

[Causal Discovery from Soft Interventions with Unknown Targets: Characterization and Learning](#)

- Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, Elias Bareinboim
- abstract@[open-review](#): One fundamental problem in the empirical sciences is of reconstructing the causal structure that underlies a phenomenon of interest through observation and experimentation. While there exists a plethora of methods capable of learning the equivalence class of causal structures that are compatible with observations, it is less well-understood how to systematically combine observations and experiments to reconstruct the underlying structure. In this paper, we investigate the task of structural learning in non-Markovian systems (i.e., when latent variables affect more than one observable) from a combination of observational and soft experimental data when the interventional targets are unknown. Using causal invariances found across the collection of observational and interventional distributions (not only conditional independencies), we define a property called psi-Markov that connects these distributions to a pair consisting of (1) a causal graph D and (2) a set of interventional targets I . Building on this property, our main contributions are two-fold: First, we provide a graphical characterization that allows one to test whether two causal graphs with possibly different sets of interventional targets belong to the same psi-Markov equivalence class. Second, we develop an algorithm capable of harnessing the collection of data to learn the corresponding equivalence class. We then prove that this algorithm is sound and complete, in the sense that it is the most informative in the sample limit, i.e., it discovers as many tails and arrowheads as can be oriented within a psi-Markov equivalence class.

[Exploiting the Surrogate Gap in Online Multiclass Classification](#)

- Dirk van der Hoeven
- abstract@[open-review](#): We present \textproc{Gaptron}, a randomized first-order algorithm for online multiclass classification. In the full information setting we provide expected mistake bounds for \textproc{Gaptron} with respect to the logistic loss, hinge loss, and the smooth hinge loss with $\$O(K\$$ regret, where the expectation is with respect to the learner's randomness and $\$K\$$ is the number of classes. In the bandit classification setting we show that \textproc{Gaptron} is the first linear time algorithm with $\$O(K\sqrt{T})\$$ expected regret. Additionally, the expected mistake bound of \textproc{Gaptron} does not depend on the dimension of the feature vector, contrary to previous algorithms with $\$O(K\sqrt{T})\$$ regret in the bandit classification setting. We present a new proof technique that exploits the gap between the zero-one loss and surrogate losses rather than exploiting properties such as exp-concavity or mixability, which are traditionally used to prove logarithmic or constant regret bounds.

[The Pitfalls of Simplicity Bias in Neural Networks](#)

- Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, Praneeth Netrapalli
- abstract@[open-review](#): Several works have proposed Simplicity Bias (SB)---the tendency of standard training procedures such as Stochastic Gradient Descent (SGD) to find simple models---to justify why neural networks generalize well [Arpit et al. 2017, Nakkiran et al. 2019, Valle-Perez et al. 2019].

However, the precise notion of simplicity remains vague. Furthermore, previous settings [Soudry et al. 2018, Gunasekar et al. 2018] that use SB to theoretically justify why neural networks generalize well do not simultaneously capture the non-robustness of neural networks---a widely observed phenomenon in practice [Goodfellow et al. 2014, Jo and Bengio 2017]. We attempt to reconcile SB and the superior standard generalization of neural networks with the non-robustness observed in practice by introducing piecewise-linear and image-based datasets, which (a) incorporate a precise notion of simplicity, (b) comprise multiple predictive features with varying levels of simplicity, and (c) capture the non-robustness of neural networks trained on real data. Using theory and empirics on these datasets, we make four observations: (i) SB of SGD and variants can be extreme: neural networks can exclusively rely on the simplest feature and remain invariant to all predictive complex features. (ii) The extreme aspect of SB could explain why seemingly benign distribution shifts and small adversarial perturbations significantly degrade model performance. (iii) Contrary to conventional wisdom, SB can also hurt generalization on the same data distribution, as SB persists even when the simplest feature has less predictive power than the more complex features. (iv) Common approaches to improve generalization and robustness---ensembles and adversarial training---can fail in mitigating SB and its pitfalls. Given the role of SB in training neural networks, we hope that the proposed datasets and methods serve as an effective testbed to evaluate novel algorithmic approaches aimed at avoiding the pitfalls of SB.

[Automatically Learning Compact Quality-aware Surrogates for Optimization Problems](#)

- Kai Wang, Bryan Wilder, Andrew Perrault, Milind Tambe
- abstract@[open-review](#): Solving optimization problems with unknown parameters often requires learning a predictive model to predict the values of the unknown parameters and then solving the problem using these values. Recent work has shown that including the optimization problem as a layer in the model training pipeline results in predictions of the unobserved parameters that lead to higher decision quality. Unfortunately, this process comes at a large computational cost because the optimization problem must be solved and differentiated through in each training iteration; furthermore, it may also sometimes fail to improve solution quality due to non-smoothness issues that arise when training through a complex optimization layer. To address these shortcomings, we learn a low-dimensional surrogate model of a large optimization problem by representing the feasible space in terms of meta-variables, each of which is a linear combination of the original variables. By training a low-dimensional surrogate model end-to-end, and jointly with the predictive model, we achieve: i) a large reduction in training and inference time; and ii) improved performance by focusing attention on the more important variables in the optimization and learning in a smoother space. Empirically, we demonstrate these improvements on a non-convex adversary modeling task, a submodular recommendation task and a convex portfolio optimization task.

[Empirical Likelihood for Contextual Bandits](#)

- Nikos Karampatziakis, John Langford, Paul Mineiro
- abstract@[open-review](#): We propose an estimator and confidence interval for computing the value of a policy from off-policy data in the contextual bandit setting. To this end we apply empirical likelihood techniques to formulate our estimator and confidence interval as simple convex optimization problems. Using the lower bound of our confidence interval, we then propose an off-policy policy optimization algorithm that searches for policies with large reward lower bound. We empirically find that both our estimator and confidence interval improve over previous proposals in finite sample regimes. Finally, the policy optimization algorithm we propose outperforms a strong baseline system for learning from off-policy data.

[Can Q-Learning with Graph Networks Learn a Generalizable Branching Heuristic for a SAT Solver?](#)

- Vitaly Kurin, Saad Godil, Shimon Whiteson, Bryan Catanzaro
- abstract@[open-review](#): We present Graph-Q-SAT, a branching heuristic for a Boolean SAT solver trained with value-based reinforcement learning (RL) using Graph Neural Networks for function approximation. Solvers using Graph-Q-SAT are complete SAT solvers that either provide a satisfying assignment or proof of unsatisfiability, which is required for many SAT applications. The branching heuristics commonly used in SAT solvers make poor decisions during their warm-up period, whereas Graph-Q-SAT is trained to examine the structure of the particular problem instance to make better decisions early in the search. Training Graph-Q-SAT is data efficient and does not require elaborate dataset preparation or feature engineering. We train Graph-Q-SAT using RL interfacing with MiniSat solver and show that Graph-Q-SAT can reduce the number of iterations required to solve SAT problems by 2-3X. Furthermore, it generalizes to unsatisfiable SAT instances, as well as to problems with 5X more variables than it was trained on. We show that for larger problems, reductions in the number of iterations lead to wall clock time reductions, the ultimate goal when designing heuristics. We also show positive zero-shot transfer behavior when testing Graph-Q-SAT on a task family different from that used for training. While more work is needed to apply Graph-Q-SAT to reduce wall clock time in modern SAT solving settings, it is a compelling proof-of-concept showing that RL equipped with Graph Neural Networks can learn a generalizable branching heuristic for SAT search.

[Non-reversible Gaussian processes for identifying latent dynamical structure in neural data](#)

- Virginia Rutten, Alberto Bernacchia, Maneesh Sahani, Guillaume Hennequin
- abstract@[open-review](#): A common goal in the analysis of neural data is to compress large population recordings into sets of interpretable, low-dimensional latent trajectories. This problem can be approached using Gaussian process (GP)-based methods which provide uncertainty quantification and principled model selection. However, standard GP priors do not distinguish between underlying dynamical processes and other forms of temporal autocorrelation. Here, we propose a new family of “dynamical” priors over trajectories, in the form of GP covariance functions that express a property shared by most dynamical systems: temporal non-reversibility. Non-reversibility is a universal signature of autonomous dynamical systems whose state trajectories follow consistent flow fields, such that any observed trajectory could not occur in reverse. Our new multi-output GP kernels can be used as drop-in replacements for standard kernels in multivariate regression, but also in latent variable models such as Gaussian process factor analysis (GPFA). We therefore introduce GPFADS (Gaussian Process Factor Analysis with Dynamical Structure), which models single-trial neural population activity using low-dimensional, non-reversible latent processes. Unlike previously proposed non-reversible multi-output kernels, ours admits a Kronecker factorization enabling fast and memory-efficient learning and inference. We apply GPFADS to synthetic data and show that it correctly recovers ground truth phase portraits. GPFADS also provides a probabilistic generalization of jPCA, a method originally developed for identifying latent rotational dynamics in neural data. When applied to monkey M1 neural recordings, GPFADS discovers latent trajectories with strong dynamical structure in the form of rotations.

[Listening to Sounds of Silence for Speech Denoising](#)

- Ruilin Xu, Rundi Wu, Yuko Ishiwaka, Carl Vondrick, Changxi Zheng
- abstract@[open-review](#): We introduce a deep learning model for speech denoising, a long-standing challenge in audio analysis arising in numerous applications. Our approach is based on a key observation about human speech: there is often a short pause between each sentence or word. In a recorded speech signal, those pauses introduce a series of time periods during which only noise is present. We leverage these incidental silent intervals to learn a model for automatic speech denoising given only mono-channel audio. Detected silent intervals over time expose not just pure noise but its time-varying features, allowing the model to learn noise dynamics and suppress it from the speech signal. Experiments on multiple datasets confirm the pivotal role of silent interval detection for speech denoising, and our method outperforms several state-of-the-art denoising methods, including those that accept only audio input (like ours) and those that denoise based on audiovisual input (and hence require more information). We also show that our method enjoys excellent generalization properties, such as denoising spoken languages not seen during training.

[BoxE: A Box Embedding Model for Knowledge Base Completion](#)

- Ralph Abboud, Ismail Ceylan, Thomas Lukasiewicz, Tommaso Salvatori
- abstract@[open-review](#): Knowledge base completion (KBC) aims to automatically infer missing facts by exploiting information already present in a knowledge base (KB). A promising approach for KBC is to embed knowledge into latent spaces and make predictions from learned embeddings. However, existing embedding models are subject to at least one of the following limitations: (1) theoretical inexpressivity, (2) lack of support for prominent inference patterns (e.g., hierarchies), (3) lack of support for KBC over higher-arity relations, and (4) lack of support for incorporating logical rules. Here, we propose a spatio-translational embedding model, called BoxE, that simultaneously addresses all these limitations. BoxE embeds entities as points, and relations as a set of hyper-rectangles (or boxes), which spatially characterize basic logical properties. This seemingly simple abstraction yields a fully expressive model offering a natural encoding for many desired logical properties. BoxE can both capture and inject rules from rich classes of rule languages, going well beyond individual inference patterns. By design, BoxE naturally applies to higher-arity KBs. We conduct a detailed experimental analysis, and show that BoxE achieves state-of-the-art performance, both on benchmark knowledge graphs and on more general KBs, and we empirically show the power of integrating logical rules.

[Coherent Hierarchical Multi-Label Classification Networks](#)

- Eleonora Giunchiglia, Thomas Lukasiewicz
- abstract@[open-review](#): Hierarchical multi-label classification (HMC) is a challenging classification task extending standard multi-label classification problems by imposing a hierarchy constraint on the classes. In this paper, we propose C-HMCNN(h), a novel approach for HMC problems, which, given a network h for the underlying multi-label classification problem, exploits the hierarchy information in order to produce predictions coherent with the constraint and improve performance. We conduct an extensive experimental analysis showing the superior performance of C-HMCNN(h) when compared to state-of-the-art models.

[Walsh-Hadamard Variational Inference for Bayesian Deep Learning](#)

- Simone Rossi, Sebastien Marmin, Maurizio Filippone
- abstract@[open-review](#): Over-parameterized models, such as DeepNets and ConvNets, form a class of models that are routinely adopted in a wide variety of applications, and for which Bayesian inference is desirable but extremely challenging. Variational inference offers the tools to tackle this challenge in a scalable way and with some degree of flexibility on the approximation, but for overparameterized models this is challenging due to the over-regularization property of the variational objective. Inspired by the literature on kernel methods, and in particular on structured approximations of distributions of random matrices, this paper proposes Walsh-Hadamard Variational Inference (WHVI), which uses Walsh-Hadamardbased factorization strategies to reduce the parameterization and accelerate computations, thus avoiding over-regularization issues with the variational objective. Extensive theoretical and empirical analyses demonstrate that WHVI yields considerable speedups and model reductions compared to other techniques to carry out approximate inference for over-parameterized models, and ultimately show how advances in kernel methods can be translated into advances in approximate Bayesian inference for Deep Learning.

[Federated Bayesian Optimization via Thompson Sampling](#)

- Zhongxiang Dai, Bryan Kian Hsiang Low, Patrick Jaillet
- abstract@[open-review](#): Bayesian optimization (BO) is a prominent approach to optimizing expensive-to-evaluate black-box functions. The massive computational capability of edge devices such as mobile phones, coupled with privacy concerns, has led to a surging interest in federated learning (FL) which focuses on collaborative training of deep neural networks (DNNs) via first-order optimization techniques. However, some common machine learning tasks such as hyperparameter tuning of DNNs lack access to gradients and thus require zeroth-order/black-box optimization. This hints at the possibility of extending BO to the FL setting (FBO) for agents to collaborate in these black-box optimization tasks. This paper presents federated Thompson sampling (FTS) which overcomes a number of key challenges of FBO and FL in a principled way: We (a) use random Fourier features to approximate the Gaussian process surrogate model used in BO, which naturally produces the parameters to be exchanged between agents, (b) design FTS based on Thompson sampling, which significantly reduces the number of parameters to be exchanged, and (c) provide a theoretical convergence guarantee that is robust against heterogeneous agents, which is a major challenge in FL and FBO. We empirically demonstrate the effectiveness of FTS in terms of communication efficiency, computational efficiency, and practical performance.

[MultiON: Benchmarking Semantic Map Memory using Multi-Object Navigation](#)

- Saim Wani, Shivansh Patel, Unnat Jain, Angel Chang, Manolis Savva
- abstract@[open-review](#): Navigation tasks in photorealistic 3D environments are challenging because they require perception and effective planning under partial observability. Recent work shows that map-like memory is useful for long-horizon navigation tasks. However, a focused investigation of the impact of maps on navigation tasks of varying complexity has not yet been performed. We propose the multiON task, which requires navigation to an episode-specific sequence of objects in a realistic environment. MultiON generalizes the ObjectGoal navigation task and explicitly tests the ability of navigation agents to locate previously observed goal objects. We perform a set of multiON experiments to examine how a variety of agent models perform across a spectrum of navigation task complexities. Our experiments show that: i) navigation performance degrades dramatically with escalating task complexity; ii) a simple semantic map agent performs surprisingly well relative to more complex neural image feature map agents; and iii) even oracle map agents achieve relatively low performance, indicating the potential for future work in training embodied navigation agents using maps.

[Neural Complexity Measures](#)

- Yoonho Lee, Juho Lee, Sung Ju Hwang, Eunho Yang, Seungjin Choi
- abstract@[open-review](#): While various complexity measures for deep neural networks exist, specifying an appropriate measure capable of predicting and explaining generalization in deep networks has proven challenging. We propose Neural Complexity (NC), a meta-learning framework for predicting generalization. Our model learns a scalar complexity measure through interactions with many heterogeneous tasks in a data-driven way. The trained NC model can be added to the standard training loss to regularize any task learner in a standard supervised learning scenario. We contrast NC's approach against existing manually-designed complexity measures and other meta-learning models, and we validate NC's performance on multiple regression and classification tasks.

[Optimal Iterative Sketching Methods with the Subsampled Randomized Hadamard Transform](#)

- Jonathan Lacotte, Sifan Liu, Edgar Dobriban, Mert Pilanci
- abstract@[open-review](#): Random projections or sketching are widely used in many algorithmic and learning contexts. Here we study the performance of iterative Hessian sketch for least-squares problems. By leveraging and extending recent results from random matrix theory on the limiting spectrum of matrices randomly projected with the subsampled randomized Hadamard transform, and truncated Haar matrices, we can study and compare the resulting algorithms to a level of precision that has not been possible before. Our technical contributions include a novel formula for the second moment of the inverse of projected matrices. We also find simple closed-form expressions for asymptotically optimal step-sizes and convergence rates. These show that the convergence rate for Haar and randomized Hadamard matrices are identical, and asymptotically improve upon Gaussian random projections. These techniques may be applied to other algorithms that employ randomized dimension reduction.

[Provably adaptive reinforcement learning in metric spaces](#)

- Tongyi Cao, Akshay Krishnamurthy
- abstract@[open-review](#): We study reinforcement learning in continuous state and action spaces endowed with a metric. We provide a refined analysis of the algorithm of Sinclair, Banerjee, and Yu (2019) and show that its regret scales with the zooming dimension of the instance. This parameter, which originates in the bandit literature, captures the size of the subsets of near optimal actions and is always smaller than the covering dimension used in previous analyses. As such, our results are the first provably adaptive guarantees for reinforcement learning in metric spaces.

[ShapeFlow: Learnable Deformation Flows Among 3D Shapes](#)

- Chiyu Jiang, Jingwei Huang, Andrea Tagliasacchi, Leonidas J. Guibas
- abstract@[open-review](#): We present ShapeFlow, a flow-based model for learning a deformation space for entire classes of 3D shapes with large intra-class variations. ShapeFlow allows learning a multi-template deformation space that is agnostic to shape topology, yet preserves fine geometric details. Different from a generative space where a latent vector is directly decoded into a shape, a deformation space decodes a vector into a continuous flow that can advect a source shape towards a target. Such a space naturally allows the disentanglement of geometric style (coming from the source) and structural pose (conforming to the target). We parametrize the deformation between geometries as a learned continuous flow field via a neural network and show that such deformations can be guaranteed to have desirable properties, such as bijectivity, freedom from self-intersections, or volume preservation. We illustrate the effectiveness of this learned deformation space for various downstream applications, including shape generation via deformation, geometric style transfer, unsupervised learning of a consistent parameterization for entire classes of shapes, and shape interpolation.

[Self-Supervised Learning by Cross-Modal Audio-Video Clustering](#)

- Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, Du Tran
- abstract@[open-review](#): Visual and audio modalities are highly correlated, yet they contain different information. Their strong correlation makes it possible to predict the semantics of one from the other with good accuracy. Their intrinsic differences make cross-modal prediction a potentially more rewarding pretext task for self-supervised learning of video and audio representations compared to within-modality learning. Based on this intuition, we propose Cross-Modal Deep Clustering (XDC), a novel self-supervised method that leverages unsupervised clustering in one modality (e.g., audio) as a supervisory signal for the other modality (e.g., video). This cross-modal supervision helps XDC utilize the semantic correlation and the differences between the two modalities. Our experiments show that XDC outperforms single-modality clustering and other multi-modal variants. XDC achieves state-of-the-art accuracy among self-supervised methods on multiple video and audio benchmarks. Most importantly, our video model pretrained on large-scale unlabeled data significantly outperforms the same model pretrained with full-supervision on ImageNet and Kinetics for action recognition on HMDB51 and UCF101. To the best of our knowledge, XDC is the first self-supervised learning method that outperforms large-scale fully-supervised pretraining for action recognition on the same architecture.

[Optimal Query Complexity of Secure Stochastic Convex Optimization](#)

- Wei Tang, Chien-Ju Ho, Yang Liu
- abstract@[open-review](#): We study the \emph{secure} stochastic convex optimization problem: a learner aims to learn the optimal point of a convex function through sequentially querying a (stochastic) gradient oracle, in the meantime, there exists an adversary who aims to free-ride and infer the learning outcome of the learner from observing the learner's queries. The adversary observes only the points of the queries but not the feedback from the oracle. The goal of the learner is to optimize the accuracy, i.e., obtaining an accurate estimate of the optimal point, while securing her privacy, i.e., making it difficult for the adversary to infer the optimal point. We formally quantify this tradeoff between learner's accuracy and privacy and characterize the lower and upper bounds on the learner's query complexity as a function of desired levels of accuracy and privacy. For the analysis of lower bounds, we provide a general template based on information theoretical analysis and then tailor the template to several families of problems, including stochastic convex optimization and (noisy) binary search. We also present a generic secure learning protocol that achieves the matching upper bound up to logarithmic factors.

[DynaBERT: Dynamic BERT with Adaptive Width and Depth](#)

- Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, Qun Liu
- abstract@[open-review](#): The pre-trained language models like BERT, though powerful in many natural language processing tasks, are both computation and memory expensive. To alleviate this problem, one approach is to compress them for specific tasks before deployment. However, recent works on BERT compression usually compress the large BERT model to a fixed smaller size, and can not fully satisfy the requirements of different edge devices with various hardware performances. In this paper, we propose a novel dynamic BERT model (abbreviated as DynaBERT), which can flexibly adjust the size and latency by selecting adaptive width and depth. The training process of DynaBERT includes first training a width-adaptive BERT and then allowing both adaptive width and depth, by distilling knowledge from the full-sized model to small sub-networks. Network rewiring is also used to keep the more important attention heads and neurons shared by more sub-networks. Comprehensive experiments under various efficiency constraints demonstrate that our proposed dynamic BERT (or RoBERTa) at its largest size has comparable performance as BERT-base (or RoBERTa-base), while at smaller widths and depths consistently outperforms existing BERT compression methods. Code is available at <https://github.com/huawei-noah/Pretrained-Language-Model/tree/master/DynaBERT>.

[Generalization Bound of Gradient Descent for Non-Convex Metric Learning](#)

- MINGZHI DONG, Xiaochen Yang, Rui Zhu, Yujiang Wang, Jing-Hao Xue
- abstract@[open-review](#): Metric learning aims to learn a distance measure that can benefit distance-based methods such as the nearest neighbour (NN) classifier. While considerable efforts have been made to improve its empirical performance and analyze its generalization ability by focusing on the data structure and model complexity, an unresolved question is how choices of algorithmic parameters, such as the number of training iterations, affect metric learning as it is typically formulated as an optimization problem and nowadays more often as a non-convex problem. In this paper, we theoretically address this question and prove the agnostic Probably Approximately Correct (PAC) learnability for metric learning algorithms with non-convex objective functions optimized via gradient descent (GD); in particular, our theoretical guarantee takes the iteration number into account. We first show that the generalization PAC bound is a sufficient condition for agnostic PAC learnability and this bound can be obtained by ensuring the uniform convergence on a densely concentrated subset of the parameter space. We then show that, for classifiers optimized via GD, their generalizability can be guaranteed if the classifier and loss function are both Lipschitz smooth, and further improved by using fewer iterations. To illustrate and exploit the theoretical findings, we finally propose a novel metric learning method called Smooth Metric and representative Instance LEarning (SMILE), designed to satisfy the Lipschitz smoothness property and learned via GD with an early stopping mechanism for better discriminability and less computational cost of NN.

[Dynamic Submodular Maximization](#)

- Morteza Monemizadeh
- abstract@[open-review](#): One of the basic primitives in the class of submodular optimization problems is the submodular maximization under a cardinality constraint. Here we are given a ground set V that is endowed with a monotone submodular function $f: 2^V \rightarrow \mathbb{R}^+$ and a parameter k

$< k \leq n$ and the goal is to return an optimal set $S \subseteq V$ of at most k elements, i.e., $f(S)$ is maximum among all subsets of V of size at most k . This basic primitive has many applications in machine learning as well as combinatorial optimization. Example applications are agglomerative clustering, exemplar-based clustering, categorical feature compression, document and corpus summarization, recommender systems, search result diversification, data subset selection, minimum spanning tree, max flow, global minimum cut, maximum matching, traveling salesman problem, max clique, max cut, set cover and knapsack, among the others. In this paper, we propose the first dynamic algorithm for this problem. Given a stream of inserts and deletes of elements of an underlying ground set V , we develop a dynamic algorithm that with high probability, maintains a $(\frac{1}{2} - \epsilon)$ -approximation of a cardinality-constrained monotone submodular maximization for any sequence of z updates (inserts and deletes) in time $O(k^2 \epsilon^{-3} \cdot \log^5 n)$, where n is the maximum size of V at any time. That is, the amortized update time of our algorithm is $O(k^2 \epsilon^{-3} \cdot \log^5 n)$.

Inference for Batched Bandits

- Kelly Zhang, Lucas Janson, Susan Murphy
- abstract@[open-review](#): As bandit algorithms are increasingly utilized in scientific studies and industrial applications, there is an associated increasing need for reliable inference methods based on the resulting adaptively-collected data. In this work, we develop methods for inference on data collected in batches using a bandit algorithm. We prove that the bandit arm selection probabilities cannot generally be assumed to concentrate. Non-concentration of the arm selection probabilities makes inference on adaptively-collected data challenging because classical statistical inference approaches, such as using asymptotic normality or the bootstrap, can have inflated Type-1 error and confidence intervals with below-nominal coverage probabilities even asymptotically. In response we develop the Batched Ordinary Least Squares estimator (BOLS) that we prove is (1) asymptotically normal on data collected from both multi-arm and contextual bandits and (2) robust to non-stationarity in the baseline reward and thus leads to reliable Type-1 error control and accurate confidence intervals.

Approximate Cross-Validation with Low-Rank Data in High Dimensions

- Will Stephenson, Madeleine Udell, Tamara Broderick
- abstract@[open-review](#): Many recent advances in machine learning are driven by a challenging trifecta: large data size N , high dimensions, and expensive algorithms. In this setting, cross-validation (CV) serves as an important tool for model assessment. Recent advances in approximate cross validation (ACV) provide accurate approximations to CV with only a single model fit, avoiding traditional CV's requirement for repeated runs of expensive algorithms. Unfortunately, these ACV methods can lose both speed and accuracy in high dimensions --- unless sparsity structure is present in the data. Fortunately, there is an alternative type of simplifying structure that is present in most data: approximate low rank (ALR). Guided by this observation, we develop a new algorithm for ACV that is fast and accurate in the presence of ALR data. Our first key insight is that the Hessian matrix --- whose inverse forms the computational bottleneck of existing ACV methods --- is ALR. We show that, despite our use of the `inverse` Hessian, a low-rank approximation using the largest (rather than the smallest) matrix eigenvalues enables fast, reliable ACV. Our second key insight is that, in the presence of ALR data, error in existing ACV methods roughly grows with the (approximate, low) rank rather than with the (full, high) dimension. These insights allow us to prove theoretical guarantees on the quality of our proposed algorithm --- along with fast-to-compute upper bounds on its error. We demonstrate the speed and accuracy of our method, as well as the usefulness of our bounds, on a range of real and simulated data sets.

GANSpace: Discovering Interpretable GAN Controls

- Erik Härkämänen, Aaron Hertzmann, Jaakko Lehtinen, Sylvain Paris
- abstract@[open-review](#): This paper describes a simple technique to analyze Generative Adversarial Networks (GANs) and create interpretable controls for image synthesis, such as change of viewpoint, aging, lighting, and time of day. We identify important latent directions based on Principal Component Analysis (PCA) applied either in latent space or feature space. Then, we show that a large number of interpretable controls can be defined by layer-wise perturbation along the principal directions. Moreover, we show that BigGAN can be controlled with layer-wise inputs in a StyleGAN-like manner. We show results on different GANs trained on various datasets, and demonstrate good qualitative matches to edit directions found through earlier supervised approaches.

Differentiable Expected Hypervolume Improvement for Parallel Multi-Objective Bayesian Optimization

- Samuel Daulton, Maximilian Balandat, Eytan Bakshy
- abstract@[open-review](#): In many real-world scenarios, decision makers seek to efficiently optimize multiple competing objectives in a sample-efficient fashion. Multi-objective Bayesian optimization (BO) is a common approach, but many of the best-performing acquisition functions do not have known analytic gradients and suffer from high computational overhead. We leverage recent advances in programming models and hardware acceleration for multi-objective BO using Expected Hypervolume Improvement (EHVI)---an algorithm notorious for its high computational complexity. We derive a novel formulation of q-Expected Hypervolume Improvement (qEHVI), an acquisition function that extends EHVI to the parallel, constrained evaluation setting. qEHVI is an exact computation of the joint EHVI of q new candidate points (up to Monte-Carlo (MC) integration error). Whereas previous EHVI formulations rely on gradient-free acquisition optimization or approximated gradients, we compute exact gradients of the MC estimator via auto-differentiation, thereby enabling efficient and effective optimization using first-order and quasi-second-order methods. Our empirical evaluation demonstrates that qEHVI is computationally tractable in many practical scenarios and outperforms state-of-the-art multi-objective BO algorithms at a fraction of their wall time.

Neuron-level Structured Pruning using Polarization Regularizer

- Tao Zhuang, Zhixuan Zhang, Yuheng Huang, Xiaoyi Zeng, Kai Shuang, Xiang Li
- abstract@[open-review](#): Neuron-level structured pruning is a very effective technique to reduce the computation of neural networks without compromising prediction accuracy. In previous works, structured pruning is usually achieved by imposing L1 regularization on the scaling factors of neurons, and pruning the neurons whose scaling factors are below a certain threshold. The reasoning is that neurons with smaller scaling factors have weaker influence on network output. A scaling factor close to 0 actually suppresses a neuron. However, L1 regularization lacks discrimination between neurons because it pushes all scaling factors towards 0. A more reasonable pruning method is to only suppress unimportant neurons (with 0 scaling factors) and simultaneously keep important neurons intact (with larger scaling factor). To achieve this goal, we propose a new regularizer on scaling factors, namely polarization regularizer. Theoretically, we prove that polarization regularizer pushes some scaling factors to 0 and others to a value $a > 0$. Experimentally, we show that structured pruning using polarization regularizer achieves much better results than using L1 regularizer. Experiments on CIFAR and ImageNet datasets show that polarization pruning achieves the state-of-the-art result to date.

Limits on Testing Structural Changes in Ising Models

- Aditya Gangrade, Bobak Nazer, Venkatesh Saligrama
- abstract@[open-review](#): We present novel information-theoretic limits on detecting sparse changes in Ising models, a problem that arises in many applications where network changes can occur due to some external stimuli. We show that the sample complexity for detecting sparse changes, in a minimax sense, is no better than learning the entire model even in settings with local sparsity. This is a surprising fact in light of prior work rooted in sparse recovery methods, which suggest that sample complexity in this context scales only with the number of network changes. To shed light on when change detection is easier than structured learning, we consider testing of edge deletion in forest-structured graphs, and high-temperature ferromagnets as case

studies. We show for these that testing of small changes is similarly hard, but testing of large changes is well-separated from structure learning. These results simply that testing of graphical models may not be amenable to concepts such as restricted strong convexity leveraged for sparsity pattern recovery, and algorithm development instead should be directed towards detection of large changes.

[Field-wise Learning for Multi-field Categorical Data](#)

- Zhibin Li, Jian Zhang, Yongshun Gong, Yazhou Yao, Qiang Wu
- abstract@[open-review](#): We propose a new method for learning with multi-field categorical data. Multi-field categorical data are usually collected over many heterogeneous groups. These groups can reflect in the categories under a field. The existing methods try to learn a universal model that fits all data, which is challenging and inevitably results in learning a complex model. In contrast, we propose a field-wise learning method leveraging the natural structure of data to learn simple yet efficient one-to-one field-focused models with appropriate constraints. In doing this, the models can be fitted to each category and thus can better capture the underlying differences in data. We present a model that utilizes linear models with variance and low-rank constraints, to help it generalize better and reduce the number of parameters. The model is also interpretable in a field-wise manner. As the dimensionality of multi-field categorical data can be very high, the models applied to such data are mostly over-parameterized. Our theoretical analysis can potentially explain the effect of over-parametrization on the generalization of our model. It also supports the variance constraints in the learning objective. The experiment results on two large-scale datasets show the superior performance of our model, the trend of the generalization error bound, and the interpretability of learning outcomes. Our code is available at <https://github.com/lzb5600/Field-wise-Learning>.

[Continual Learning in Low-rank Orthogonal Subspaces](#)

- Arslan Chaudhry, Naeemullah Khan, Puneet Dokania, Philip Torr
- abstract@[open-review](#): In continual learning (CL), a learner is faced with a sequence of tasks, arriving one after the other, and the goal is to remember all the tasks once the continual learning experience is finished. The prior art in CL uses episodic memory, parameter regularization or extensible network structures to reduce interference among tasks, but in the end, all the approaches learn different tasks in a joint vector space. We believe this invariably leads to interference among different tasks. We propose to learn tasks in different (low-rank) vector subspaces that are kept orthogonal to each other in order to minimize interference. Further, to keep the gradients of different tasks coming from these subspaces orthogonal to each other, we learn isometric mappings by posing network training as an optimization problem over the Stiefel manifold. To the best of our understanding, we report, for the first time, strong results over experience-replay baseline with and without memory on standard classification benchmarks in continual learning.

[Unsupervised Learning of Visual Features by Contrasting Cluster Assignments](#)

- Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, Armand Joulin
- abstract@[open-review](#): Unsupervised image representations have significantly reduced the gap with supervised pretraining, notably with the recent achievements of contrastive learning methods. These contrastive methods typically work online and rely on a large number of explicit pairwise feature comparisons, which is computationally challenging. In this paper, we propose an online algorithm, SwAV, that takes advantage of contrastive methods without requiring to compute pairwise comparisons. Specifically, our method simultaneously clusters the data while enforcing consistency between cluster assignments produced for different augmentations (or views) of the same image, instead of comparing features directly as in contrastive learning. Simply put, we use a swapped prediction mechanism where we predict the code of a view from the representation of another view. Our method can be trained with large and small batches and can scale to unlimited amounts of data. Compared to previous contrastive methods, our method is more memory efficient since it does not require a large memory bank or a special momentum network. In addition, we also propose a new data augmentation strategy, multi-crop, that uses a mix of views with different resolutions in place of two full-resolution views, without increasing the memory or compute requirements. We validate our findings by achieving 75.3% top-1 accuracy on ImageNet with ResNet-50, as well as surpassing supervised pretraining on all the considered transfer tasks.

[Sharpened Generalization Bounds based on Conditional Mutual Information and an Application to Noisy, Iterative Algorithms](#)

- Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M. Roy, Gintare Karolina Dziugaite
- abstract@[open-review](#): The information-theoretic framework of Russo and Zou (2016) and Xu and Raginsky (2017) provides bounds on the generalization error of a learning algorithm in terms of the mutual information between the algorithm's output and the training sample. In this work, we study the proposal, by Steinke and Zakynthinou (2020), to reason about the generalization error of a learning algorithm by introducing a super sample that contains the training sample as a random subset and computing mutual information conditional on the super sample. We first show that these new bounds based on the conditional mutual information are tighter than those based on the unconditional mutual information. We then introduce yet tighter bounds, building on the "individual sample" idea of Bu et al. (2019) and the "data dependent" ideas of Negrea et al. (2019), using disintegrated mutual information. Finally, we apply these bounds to the study of Langevin dynamics algorithm, showing that conditioning on the super sample allows us to exploit information in the optimization trajectory to obtain tighter bounds based on hypothesis tests.

[Learning Deformable Tetrahedral Meshes for 3D Reconstruction](#)

- Jun Gao, Wenzheng Chen, Tommy Xiang, Alec Jacobson, Morgan McGuire, Sanja Fidler
- abstract@[open-review](#): 3D shape representations that accommodate learning-based 3D reconstruction are an open problem in machine learning and computer graphics. Previous work on neural 3D reconstruction demonstrated benefits, but also limitations, of point cloud, voxel, surface mesh, and implicit function representations. We introduce \emph{Deformable Tetrahedral Meshes} (DefTet) as a particular parameterization that utilizes volumetric tetrahedral meshes for the reconstruction problem. Unlike existing volumetric approaches, DefTet optimizes for both vertex placement and occupancy, and is differentiable with respect to standard 3D reconstruction loss functions. It is thus simultaneously high-precision, volumetric, and amenable to learning-based neural architectures. We show that it can represent arbitrary, complex topology, is both memory and computationally efficient, and can produce high-fidelity reconstructions with a significantly smaller grid size than alternative volumetric approaches. The predicted surfaces are also inherently defined as tetrahedral meshes, thus do not require post-processing. We demonstrate that DefTet matches or exceeds both the quality of the previous best approaches and the performance of the fastest ones. Our approach obtains high-quality tetrahedral meshes computed directly from noisy point clouds, and is the first to showcase high-quality 3D results using only a single image as input.

[Information theoretic limits of learning a sparse rule](#)

- Clément Luneau, Jean Barbier, Nicolas Macris
- abstract@[open-review](#): We consider generalized linear models in regimes where the number of nonzero components of the signal and accessible data points are sublinear with respect to the size of the signal. We prove a variational formula for the asymptotic mutual information per sample when the system size grows to infinity. This result allows us to derive an expression for the minimum mean-square error (MMSE) of the Bayesian estimator when the signal entries have a discrete distribution with finite support. We find that, for such signals and suitable vanishing scalings of the sparsity and sampling rate, the MMSE is nonincreasing piecewise constant. In specific instances the MMSE even displays an all-or-nothing phase transition, that is, the MMSE sharply jumps from its maximum value to zero at a critical sampling rate. The all-or-nothing phenomenon has previously been shown to occur in high-dimensional linear regression. Our analysis goes beyond the linear case and applies to learning the weights of a perceptron with general activation function in a teacher-student scenario. In particular, we discuss an all-or-nothing phenomenon for the generalization error with a sublinear set of training examples.

Self-supervised learning through the eyes of a child

- Emin Orhan, Vaibhav Gupta, Brenden M. Lake
- abstract@[open-review](#): Within months of birth, children develop meaningful expectations about the world around them. How much of this early knowledge can be explained through generic learning mechanisms applied to sensory data, and how much of it requires more substantive innate inductive biases? Addressing this fundamental question in its full generality is currently infeasible, but we can hope to make real progress in more narrowly defined domains, such as the development of high-level visual categories, thanks to improvements in data collecting technology and recent progress in deep learning. In this paper, our goal is precisely to achieve such progress by utilizing modern self-supervised deep learning methods and a recent longitudinal, egocentric video dataset recorded from the perspective of three young children (Sullivan et al., 2020). Our results demonstrate the emergence of powerful, high-level visual representations from developmentally realistic natural videos using generic self-supervised learning objectives.

Unsupervised Semantic Aggregation and Deformable Template Matching for Semi-Supervised Learning

- Tao Han, Junyu Gao, Yuan Yuan, Qi Wang
- abstract@[open-review](#): Unlabeled data learning has attracted considerable attention recently. However, it is still elusive to extract the expected high-level semantic feature with mere unsupervised learning. In the meantime, semi-supervised learning (SSL) demonstrates a promising future in leveraging few samples. In this paper, we combine both to propose an Unsupervised Semantic Aggregation and Deformable Template Matching (USADTM) framework for SSL, which strives to improve the classification performance with few labeled data and then reduce the cost in data annotating. Specifically, unsupervised semantic aggregation based on Triplet Mutual Information (T-MI) loss is explored to generate semantic labels for unlabeled data. Then the semantic labels are aligned to the actual class by the supervision of labeled data. Furthermore, a feature pool that stores the labeled samples is dynamically updated to assign proxy labels for unlabeled data, which are used as targets for cross-entropy minimization. Extensive experiments and analysis across four standard semi-supervised learning benchmarks validate that USADTM achieves top performance (e.g., 90.46% accuracy on CIFAR-10 with 40 labels and 95.20% accuracy with 250 labels). The code is released at <https://github.com/taohan10200/USADTM>.

A game-theoretic analysis of networked system control for common-pool resource management using multi-agent reinforcement learning

- Arnu Pretorius, Scott Cameron, Elan van Biljon, Thomas Makkink, Shahil Mawjee, Jeremy du Plessis, Jonathan Shock, Alexandre Laterre, Karim Beguir
- abstract@[open-review](#): Multi-agent reinforcement learning has recently shown great promise as an approach to networked system control. Arguably, one of the most difficult and important tasks for which large scale networked system control is applicable is common-pool resource management. Crucial common-pool resources include arable land, fresh water, wetlands, wildlife, fish stock, forests and the atmosphere, of which proper management is related to some of society's greatest challenges such as food security, inequality and climate change. Here we take inspiration from a recent research program investigating the game-theoretic incentives of humans in social dilemma situations such as the well-known \textit{tragedy of the commons}. However, instead of focusing on biologically evolved human-like agents, our concern is rather to better understand the learning and operating behaviour of engineered networked systems comprising general-purpose reinforcement learning agents, subject only to nonbiological constraints such as memory, computation and communication bandwidth. Harnessing tools from empirical game-theoretic analysis, we analyse the differences in resulting solution concepts that stem from employing different information structures in the design of networked multi-agent systems. These information structures pertain to the type of information shared between agents as well as the employed communication protocol and network topology. Our analysis contributes new insights into the consequences associated with certain design choices and provides an additional dimension of comparison between systems beyond efficiency, robustness, scalability and mean control performance.

What shapes feature representations? Exploring datasets, architectures, and training

- Katherine Hermann, Andrew Lampinen
- abstract@[open-review](#): In naturalistic learning problems, a model's input contains a wide range of features, some useful for the task at hand, and others not. Of the useful features, which ones does the model use? Of the task-irrelevant features, which ones does the model represent? Answers to these questions are important for understanding the basis of models' decisions, as well as for building models that learn versatile, adaptable representations useful beyond the original training task. We study these questions using synthetic datasets in which the task-relevance of input features can be controlled directly. We find that when two features redundantly predict the labels, the model preferentially represents one, and its preference reflects what was most linearly decodable from the untrained model. Over training, task-relevant features are enhanced, and task-irrelevant features are partially suppressed. Interestingly, in some cases, an easier, weakly predictive feature can suppress a more strongly predictive, but more difficult one. Additionally, models trained to recognize both easy and hard features learn representations most similar to models that use only the easy feature. Further, easy features lead to more consistent representations across model runs than do hard features. Finally, models have greater representational similarity to an untrained model than to models trained on a different task. Our results highlight the complex processes that determine which features a model represents.

Optimal Best-arm Identification in Linear Bandits

- Yassir Jedra, Alexandre Proutiere
- abstract@[open-review](#): We study the problem of best-arm identification with fixed confidence in stochastic linear bandits. The objective is to identify the best arm with a given level of certainty while minimizing the sampling budget. We devise a simple algorithm whose sampling complexity matches known instance-specific lower bounds, asymptotically almost surely and in expectation. The algorithm relies on an arm sampling rule that tracks an optimal proportion of arm draws, and that remarkably can be updated as rarely as we wish, without compromising its theoretical guarantees. Moreover, unlike existing best-arm identification strategies, our algorithm uses a stopping rule that does not depend on the number of arms. Experimental results suggest that our algorithm significantly outperforms existing algorithms. The paper further provides a first analysis of the best-arm identification problem in linear bandits with a continuous set of arms.

Data Diversification: A Simple Strategy For Neural Machine Translation

- Xuan-Phi Nguyen, Shafiq Joty, Kui Wu, Ai Ti Aw
- abstract@[open-review](#): We introduce Data Diversification: a simple but effective strategy to boost neural machine translation (NMT) performance. It diversifies the training data by using the predictions of multiple forward and backward models and then merging them with the original dataset on which the final NMT model is trained. Our method is applicable to all NMT models. It does not require extra monolingual data like back-translation, nor does it add more computations and parameters like ensembles of models. Our method achieves state-of-the-art BLEU scores of 30.7 and 43.7 in the WMT14 English-German and English-French translation tasks, respectively. It also substantially improves on 8 other translation tasks: 4 IWSLT tasks (English-German and English-French) and 4 low-resource translation tasks (English-Nepali and English-Sinhala). We demonstrate that our method is more effective than knowledge distillation and dual learning, it exhibits strong correlation with ensembles of models, and it trades perplexity off for better BLEU score.

Interstellar: Searching Recurrent Architecture for Knowledge Graph Embedding

- Yongqi Zhang, Quanming Yao, Lei Chen

- abstract@[open-review](#): Knowledge graph (KG) embedding is well-known in learning representations of KGs. Many models have been proposed to learn the interactions between entities and relations of the triplets. However, long-term information among multiple triplets is also important to KG. In this work, based on the relational paths, which are composed of a sequence of triplets, we define the Interstellar as a recurrent neural architecture search problem for the short-term and long-term information along the paths. First, we analyze the difficulty of using a unified model to work as the Interstellar. Then, we propose to search for recurrent architecture as the Interstellar for different KG tasks. A case study on synthetic data illustrates the importance of the defined search problem. Experiments on real datasets demonstrate the effectiveness of the searched models and the efficiency of the proposed hybrid-search algorithm.

[CoSE: Compositional Stroke Embeddings](#)

- Emre Aksan, Thomas Deselaers, Andrea Tagliasacchi, Otmar Hilliges
- abstract@[open-review](#): We present a generative model for stroke-based drawing tasks which is able to model complex free-form structures. While previous approaches rely on sequence-based models for drawings of basic objects or handwritten text, we propose a model that treats drawings as a collection of strokes that can be composed into complex structures such as diagrams (e.g., flow-charts). At the core of the approach lies a novel auto-encoder that projects variable-length strokes into a latent space of fixed dimension. This representation space allows a relational model, operating in latent space, to better capture the relationship between strokes and to predict subsequent strokes. We demonstrate qualitatively and quantitatively that our proposed approach is able to model the appearance of individual strokes, as well as the compositional structure of larger diagram drawings. Our approach is suitable for interactive use cases such as auto-completing diagrams. We make code and models publicly available at <https://eth-ait.github.io/pose>.

[Learning Multi-Agent Coordination for Enhancing Target Coverage in Directional Sensor Networks](#)

- Jing Xu, Fangwei Zhong, Yizhou Wang
- abstract@[open-review](#): Maximum target coverage by adjusting the orientation of distributed sensors is an important problem in directional sensor networks (DSNs). This problem is challenging as the targets usually move randomly but the coverage range of sensors is limited in angle and distance. Thus, it is required to coordinate sensors to get ideal target coverage with low power consumption, e.g. no missing targets or reducing redundant coverage. To realize this, we propose a Hierarchical Target-oriented Multi-Agent Coordination (HiT-MAC), which decomposes the target coverage problem into two-level tasks: targets assignment by a coordinator and tracking assigned targets by executors. Specifically, the coordinator periodically monitors the environment globally and allocates targets to each executor. In turn, the executor only needs to track its assigned targets. To effectively learn the HiT-MAC by reinforcement learning, we further introduce a bunch of practical methods, including a self-attention module, marginal contribution approximation for the coordinator, goal-conditional observation filter for the executor, etc. Empirical results demonstrate the advantage of HiT-MAC in coverage rate, learning efficiency, and scalability, comparing to baselines. We also conduct an ablative analysis on the effectiveness of the introduced components in the framework.

[Biological credit assignment through dynamic inversion of feedforward networks](#)

- Bill Podlaski, Christian K. Machens
- abstract@[open-review](#): Learning depends on changes in synaptic connections deep inside the brain. In multilayer networks, these changes are triggered by error signals fed back from the output, generally through a stepwise inversion of the feedforward processing steps. The gold standard for this process --- backpropagation --- works well in artificial neural networks, but is biologically implausible. Several recent proposals have emerged to address this problem, but many of these biologically-plausible schemes are based on learning an independent set of feedback connections. This complicates the assignment of errors to each synapse by making it dependent upon a second learning problem, and by fitting inversions rather than guaranteeing them. Here, we show that feedforward network transformations can be effectively inverted through dynamics. We derive this dynamic inversion from the perspective of feedback control, where the forward transformation is reused and dynamically interacts with fixed or random feedback to propagate error signals during the backward pass. Importantly, this scheme does not rely upon a second learning problem for feedback because accurate inversion is guaranteed through the network dynamics. We map these dynamics onto generic feedforward networks, and show that the resulting algorithm performs well on several supervised and unsupervised datasets. Finally, we discuss potential links between dynamic inversion and second-order optimization. Overall, our work introduces an alternative perspective on credit assignment in the brain, and proposes a special role for temporal dynamics and feedback control during learning.

[Discriminative Sounding Objects Localization via Self-supervised Audiovisual Matching](#)

- Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, Dejing Dou
- abstract@[open-review](#): Discriminatively localizing sounding objects in cocktail-party, i.e., mixed sound scenes, is commonplace for humans, but still challenging for machines. In this paper, we propose a two-stage learning framework to perform self-supervised class-aware sounding object localization. First, we propose to learn robust object representations by aggregating the candidate sound localization results in the single source scenes. Then, class-aware object localization maps are generated in the cocktail-party scenarios by referring the pre-learned object knowledge, and the sounding objects are accordingly selected by matching audio and visual object category distributions, where the audiovisual consistency is viewed as the self-supervised signal. Experimental results in both realistic and synthesized cocktail-party videos demonstrate that our model is superior in filtering out silent objects and pointing out the location of sounding objects of different classes. Code is available at <https://github.com/DTao/Discriminative-Sounding-Objects-Localization>.

[Learning Multi-Agent Communication through Structured Attentive Reasoning](#)

- Murtaza Rangwala, Ryan Williams
- abstract@[open-review](#): Learning communication via deep reinforcement learning has recently been shown to be an effective way to solve cooperative multi-agent tasks. However, learning which communicated information is beneficial for each agent's decision-making process remains a challenging task. In order to address this problem, we explore relational reinforcement learning which leverages attention-based networks to learn efficient and interpretable relations between entities. On the foundation of relations, we introduce a novel communication architecture that exploits a memory-based attention network that selectively reasons about the value of information received from other agents while considering its past experiences. Specifically, the model communicates by first computing the relevance of messages received from other agents and then extracts task-relevant information from memories given the newly received information. We empirically demonstrate the strength of our model in cooperative and competitive multi-agent tasks, where inter-agent communication and reasoning over prior information substantially improves performance compared to baselines. We further show in the accompanying videos and experimental results that the agents learn a sophisticated and diverse set of cooperative behaviors to solve challenging tasks, both for discrete and continuous action spaces using on-policy and off-policy gradient methods. By developing an explicit architecture that is targeted towards communication, our work aims to open new directions to overcome important challenges in multi-agent cooperation through learned communication.

[Private Identity Testing for High-Dimensional Distributions](#)

- Clément L. Canonne, Gautam Kamath, Audra McMillan, Jonathan Ullman, Lydia Zakynthinou
- abstract@[open-review](#): In this work we present novel differentially private identity (goodness-of-fit) testers for natural and widely studied classes of multivariate product distributions: Gaussians in \mathbb{R}^d with known covariance and product distributions over $\{\pm 1\}^d$. Our testers have improved sample complexity compared to those derived from previous techniques, and are the first testers whose sample complexity matches the order-optimal minimax

sample complexity of $O(d^{1/2}/\alpha^2)$ in many parameter regimes. We construct two types of testers, exhibiting tradeoffs between sample complexity and computational complexity. Finally, we provide a two-way reduction between testing a subclass of multivariate product distributions and testing univariate distributions, and thereby obtain upper and lower bounds for testing this subclass of product distributions.

[On the Optimal Weighted \$\ell_2\$ Regularization in Overparameterized Linear Regression](#)

- Denny Wu, Ji Xu
- abstract@[open-review](#): We consider the linear model $\mathbf{y} = \mathbf{X}\beta_{\star} + \epsilon$ with $\mathbf{X} \in \mathbb{R}^{n \times p}$ in the overparameterized regime $p > n$. We estimate β_{\star} via generalized (weighted) ridge regression: $\hat{\beta} = (\lambda(\mathbf{X}^\top \mathbf{X} + \lambda \Sigma_w)^{-1} \mathbf{X}^\top \mathbf{y})$, where Σ_w is the weighting matrix. Assuming a random effects model with general data covariance Σ_x and anisotropic prior on the true coefficients β_{\star} , i.e., $\mathbb{E}[\beta_{\star} \beta_{\star}^\top] = \Sigma_{\beta_{\star}}$, we provide an exact characterization of the prediction risk $\mathbb{E}[(\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta})]$ in the proportional asymptotic limit $p/n \rightarrow \gamma$. Our general setup leads to a number of interesting findings. We outline precise conditions that decide the sign of the optimal setting λ_{opt} for the ridge parameter λ and confirm the implicit regularization effect of overparameterization, which theoretically justifies the surprising empirical observation that λ_{opt} can be negative in the overparameterized regime. We also characterize the double descent phenomenon for principal component regression (PCR) when \mathbf{X} and β_{\star} are both anisotropic. Finally, we determine the optimal weighting matrix Σ_w for both the ridgeless ($\lambda \rightarrow 0$) and optimally regularized ($\lambda = \lambda_{\text{opt}}$) case, and demonstrate the advantage of the weighted objective over standard ridge regression and PCR.

[An Efficient Asynchronous Method for Integrating Evolutionary and Gradient-based Policy Search](#)

- Kyunghyun Lee, Byeong-Uk Lee, Ukcheol Shin, In So Kweon
- abstract@[open-review](#): Deep reinforcement learning (DRL) algorithms and evolution strategies (ES) have been applied to various tasks, showing excellent performances. These have the opposite properties, with DRL having good sample efficiency and poor stability, while ES being vice versa. Recently, there have been attempts to combine these algorithms, but these methods fully rely on synchronous update scheme, making it not ideal to maximize the benefits of the parallelism in ES. To solve this challenge, asynchronous update scheme was introduced, which is capable of good time-efficiency and diverse policy exploration. In this paper, we introduce an Asynchronous Evolution Strategy-Reinforcement Learning (AES-RL) that maximizes the parallel efficiency of ES and integrates it with policy gradient methods. Specifically, we propose 1) a novel framework to merge ES and DRL asynchronously and 2) various asynchronous update methods that can take all advantages of asynchronism, ES, and DRL, which are exploration and time efficiency, stability, and sample efficiency, respectively. The proposed framework and update methods are evaluated in continuous control benchmark work, showing superior performance as well as time efficiency compared to the previous methods.

[MetaSDF: Meta-Learning Signed Distance Functions](#)

- Vincent Sitzmann, Eric Chan, Richard Tucker, Noah Snavely, Gordon Wetzstein
- abstract@[open-review](#): Neural implicit shape representations are an emerging paradigm that offers many potential benefits over conventional discrete representations, including memory efficiency at a high spatial resolution. Generalizing across shapes with such neural implicit representations amounts to learning priors over the respective function space and enables geometry reconstruction from partial or noisy observations. Existing generalization methods rely on conditioning a neural network on a low-dimensional latent code that is either regressed by an encoder or jointly optimized in the auto-decoder framework. Here, we formalize learning of a shape space as a meta-learning problem and leverage gradient-based meta-learning algorithms to solve this task. We demonstrate that this approach performs on par with auto-decoder based approaches while being an order of magnitude faster at test-time inference. We further demonstrate that the proposed gradient-based method outperforms encoder-decoder based methods that leverage pooling-based set encoders.

[Simple and Scalable Sparse k-means Clustering via Feature Ranking](#)

- Zhiyue Zhang, Kenneth Lange, Jason Xu
- abstract@[open-review](#): Clustering, a fundamental activity in unsupervised learning, is notoriously difficult when the feature space is high-dimensional. Fortunately, in many realistic scenarios, only a handful of features are relevant in distinguishing clusters. This has motivated the development of sparse clustering techniques that typically rely on k-means within outer algorithms of high computational complexity. Current techniques also require careful tuning of shrinkage parameters, further limiting their scalability. In this paper, we propose a novel framework for sparse k-means clustering that is intuitive, simple to implement, and competitive with state-of-the-art algorithms. We show that our algorithm enjoys consistency and convergence guarantees. Our core method readily generalizes to several task-specific algorithms such as clustering on subsets of attributes and in partially observed data settings. We showcase these contributions thoroughly via simulated experiments and real data benchmarks, including a case study on protein expression in trisomic mice.

[Model-based Adversarial Meta-Reinforcement Learning](#)

- Zichuan Lin, Garrett Thomas, Guangwen Yang, Tengyu Ma
- abstract@[open-review](#): Meta-reinforcement learning (meta-RL) aims to learn from multiple training tasks the ability to adapt efficiently to unseen test tasks. Despite the success, existing meta-RL algorithms are known to be sensitive to the task distribution shift. When the test task distribution is different from the training task distribution, the performance may degrade significantly. To address this issue, this paper proposes Model-based Adversarial Meta-Reinforcement Learning (AdMRL), where we aim to minimize the worst-case sub-optimality gap --- the difference between the optimal return and the return that the algorithm achieves after adaptation --- across all tasks in a family of tasks, with a model-based approach. We propose a minimax objective and optimize it by alternating between learning the dynamics model on a fixed task and finding the adversarial task for the current model --- the task for which the policy induced by the model is maximally suboptimal. Assuming the family of tasks is parameterized, we derive a formula for the gradient of the suboptimality with respect to the task parameters via the implicit function theorem, and show how the gradient estimator can be efficiently implemented by the conjugate gradient method and a novel use of the REINFORCE estimator. We evaluate our approach on several continuous control benchmarks and demonstrate its efficacy in the worst-case performance over all tasks, the generalization power to out-of-distribution tasks, and in training and test time sample efficiency, over existing state-of-the-art meta-RL algorithms.

[Graph Policy Network for Transferable Active Learning on Graphs](#)

- Shengding Hu, Zheng Xiong, Meng Qu, Xingdi Yuan, Marc-Alexandre Côté, Zhiyuan Liu, Jian Tang
- abstract@[open-review](#): Graph neural networks (GNNs) have been attracting increasing popularity due to their simplicity and effectiveness in a variety of fields. However, a large number of labeled data is generally required to train these networks, which could be very expensive to obtain in some domains. In this paper, we study active learning for GNNs, i.e., how to efficiently label the nodes on a graph to reduce the annotation cost of training GNNs. We formulate the problem as a sequential decision process on graphs and train a GNN-based policy network with reinforcement learning to learn the optimal query strategy. By jointly training on several source graphs with full labels, we learn a transferable active learning policy which can directly generalize to unlabeled target graphs. Experimental results on multiple datasets from different domains prove the effectiveness of the learned policy in promoting active learning performance in both settings of transferring between graphs in the same domain and across different domains.

[Towards a Better Global Loss Landscape of GANs](#)

- Ruoyu Sun, Tiantian Fang, Alexander Schwing
- abstract@[open-review](#): Understanding of GAN training is still very limited. One major challenge is its non-convex-non-concave min-max objective, which may lead to sub-optimal local minima. In this work, we perform a global landscape analysis of the empirical loss of GANs. We prove that a class of separable-GAN, including the original JS-GAN, has exponentially many bad basins which are perceived as mode-collapse. We also study the relativistic pairing GAN (RpGAN) loss which couples the generated samples and the true samples. We prove that RpGAN has no bad basins. Experiments on synthetic data show that the predicted bad basin can indeed appear in training. We also perform experiments to support our theory that RpGAN has a better landscape than separable-GAN. For instance, we empirically show that RpGAN performs better than separable-GAN with relatively narrow neural nets. The code is available at \url{https://github.com/AilsaF/RS-GAN}.

[Weighted QMIX: Expanding Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning](#)

- Tabish Rashid, Gregory Farquhar, Bei Peng, Shimon Whiteson
- abstract@[open-review](#): QMIX is a popular \$Q\$-learning algorithm for cooperative MARL in the centralised training and decentralised execution paradigm. In order to enable easy decentralisation, QMIX restricts the joint action \$Q\$-values it can represent to be a monotonic mixing of each agent's utilities. However, this restriction prevents it from representing value functions in which an agent's ordering over its actions can depend on other agents' actions. To analyse this representational limitation, we first formalise the objective QMIX optimises, which allows us to view QMIX as an operator that first computes the \$Q\$-learning targets and then projects them into the space representable by QMIX. This projection returns a representable \$Q\$-value that minimises the unweighted squared error across all joint actions. We show in particular that this projection can fail to recover the optimal policy even with access to \$Q^\pi\$, which primarily stems from the equal weighting placed on each joint action. We rectify this by introducing a weighting into the projection, in order to place more importance on the better joint actions. We propose two weighting schemes and prove that they recover the correct maximal action for any joint action \$Q\$-values, and therefore for \$Q^\pi\$ as well. Based on our analysis and results in the tabular setting we introduce two scalable versions of our algorithm, Centrally-Weighted (CW) QMIX and Optimistically-Weighted (OW) QMIX and demonstrate improved performance on both predator-prey and challenging multi-agent StarCraft benchmark tasks (Samvelyan et al., 2019).

[BanditPAM: Almost Linear Time k-Medoids Clustering via Multi-Armed Bandits](#)

- Mo Tiwari, Martin J. Zhang, James Mayclin, Sebastian Thrun, Chris Piech, Ilan Shomorony
- abstract@[open-review](#): Clustering is a ubiquitous task in data science. Compared to the commonly used k-means clustering, k-medoids clustering requires the cluster centers to be actual data points and supports arbitrary distance metrics, which permits greater interpretability and the clustering of structured objects. Current state-of-the-art k-medoids clustering algorithms, such as Partitioning Around Medoids (PAM), are iterative and are quadratic in the dataset size n for each iteration, being prohibitively expensive for large datasets. We propose BanditPAM, a randomized algorithm inspired by techniques from multi-armed bandits, that reduces the complexity of each PAM iteration from $O(n^2)$ to $O(n\log n)$ and returns the same results with high probability, under assumptions on the data that often hold in practice. As such, BanditPAM matches state-of-the-art clustering loss while reaching solutions much faster. We empirically validate our results on several large real-world datasets, including a coding exercise submissions dataset from Code.org, the 10x Genomics 68k PBMC single-cell RNA sequencing dataset, and the MNIST handwritten digits dataset. In these experiments, we observe that BanditPAM returns the same results as state-of-the-art PAM-like algorithms up to 4x faster while performing up to 200x fewer distance computations. The improvements demonstrated by BanditPAM enable k-medoids clustering on a wide range of applications, including identifying cell types in large-scale single-cell data and providing scalable feedback for students learning computer science online. We also release highly optimized Python and C++ implementations of our algorithm.

[UDH: Universal Deep Hiding for Steganography, Watermarking, and Light Field Messaging](#)

- Chaoning Zhang, Philipp Benz, Adil Karjauv, Geng Sun, In So Kweon
- abstract@[open-review](#): Neural networks have been shown effective in deep steganography for hiding a full image in another. However, the reason for its success remains not fully clear. Under the existing cover (\$C\$) dependent deep hiding (DDH) pipeline, it is challenging to analyze how the secret (\$S\$) image is encoded since the encoded message cannot be analyzed independently. We propose a novel universal deep hiding (UDH) meta-architecture to disentangle the encoding of \$S\$ from \$C\$. We perform extensive analysis and demonstrate that the success of deep steganography can be attributed to a frequency discrepancy between \$C\$ and the encoded secret image. Despite \$S\$ being hidden in a cover-agnostic manner, strikingly, UDH achieves a performance comparable to the existing DDH. Beyond hiding one image, we push the limits of deep steganography. Exploiting its property of being universal, we propose universal watermarking as a timely solution to address the concern of the exponentially increasing amount of images/videos. UDH is robust to a pixel intensity shift on the container image, which makes it suitable for challenging application of light field messaging (LFM). This is the first work demonstrating the success of (DNN-based) hiding a full image for watermarking and LFM. Code: \url{https://github.com/ChaoningZhang/Universal-Deep-Hiding}

[Evidential Sparsification of Multimodal Latent Spaces in Conditional Variational Autoencoders](#)

- Masha Itkina, Boris Ivanovic, Ransalu Senanayake, Mykel J. Kochenderfer, Marco Pavone
- abstract@[open-review](#): Discrete latent spaces in variational autoencoders have been shown to effectively capture the data distribution for many real-world problems such as natural language understanding, human intent prediction, and visual scene representation. However, discrete latent spaces need to be sufficiently large to capture the complexities of real-world data, rendering downstream tasks computationally challenging. For instance, performing motion planning in a high-dimensional latent representation of the environment could be intractable. We consider the problem of sparsifying the discrete latent space of a trained conditional variational autoencoder, while preserving its learned multimodality. As a post hoc latent space reduction technique, we use evidential theory to identify the latent classes that receive direct evidence from a particular input condition and filter out those that do not. Experiments on diverse tasks, such as image generation and human behavior prediction, demonstrate the effectiveness of our proposed technique at reducing the discrete latent sample space size of a model while maintaining its learned multimodality.

[An Unbiased Risk Estimator for Learning with Augmented Classes](#)

- Yu-Jie Zhang, Peng Zhao, Lanjihong Ma, Zhi-Hua Zhou
- abstract@[open-review](#): This paper studies the problem of learning with augmented classes (LAC), where augmented classes unobserved in the training data might emerge in the testing phase. Previous studies generally attempt to discover augmented classes by exploiting geometric properties, achieving inspiring empirical performance yet lacking theoretical understandings particularly on the generalization ability. In this paper we show that, by using unlabeled training data to approximate the potential distribution of augmented classes, an unbiased risk estimator of the testing distribution can be established for the LAC problem under mild assumptions, which paves a way to develop a sound approach with theoretical guarantees. Moreover, the proposed approach can adapt to complex changing environments where augmented classes may appear and the prior of known classes may change simultaneously. Extensive experiments confirm the effectiveness of our proposed approach.

[AutoBSS: An Efficient Algorithm for Block Stacking Style Search](#)

- Yikang Zhang, Jian Zhang, Zhao Zhong
- abstract@[open-review](#): Neural network architecture design mostly focuses on the new convolutional operator or special topological structure of network block, little attention is drawn to the configuration of stacking each block, called Block Stacking Style (BSS). Recent studies show that BSS may also have an unnegelectable impact on networks, thus we design an efficient algorithm to search it automatically. The proposed method, AutoBSS, is a novel AutoML algorithm based on Bayesian optimization by iteratively refining and clustering Block Stacking Style Code (BSSC), which can find optimal BSS in a few trials without biased evaluation. On ImageNet classification task, ResNet50/MobileNetV2/EfficientNet-B0 with our searched BSS achieve 79.29%/74.5%/77.79%, which outperform the original baselines by a large margin. More importantly, experimental results on model compression, object detection and instance segmentation show the strong generalizability of the proposed AutoBSS, and further verify the unnegelectable impact of BSS on neural networks.

Pushing the Limits of Narrow Precision Inferencing at Cloud Scale with Microsoft Floating Point

- Bita Darvish Rouhani, Daniel Lo, Ritchie Zhao, Ming Liu, Jeremy Fowers, Kalin Ovtcharov , Anna Vinogradsky, Sarah Massengill , Lita Yang, Ray Bittner, Alessandro Forin, Haishan Zhu, Taesik Na, Prerak Patel, Shuai Che, Lok Chand Koppaka , XIA SONG, Subhojit Som, Kaustav Das, Saurabh T, Steve Reinhardt , Sitaram Lanka, Eric Chung, Doug Burger
- abstract@[open-review](#): In this paper, we explore the limits of Microsoft Floating Point (MSFP), a new class of datatypes developed for production cloud-scale inferencing on custom hardware. Through the co-evolution of hardware design and algorithms, MSFP achieves accuracy comparable to or better than industry standards Bfloat16 and INT8 at 3x and 4x lower cost, respectively. MSFP incurs negligible impact to accuracy (<1%), requires no changes to the model topology, and is integrated with a mature cloud production pipeline. MSFP supports various classes of deep learning models including CNNs, RNNs, and Transformers without modification. Finally, we characterize the accuracy and implementation of MSFP and demonstrate its efficacy on a number of production scenarios, including models that power major online scenarios such as web search, question-answering, and image classification.

Stochastic Optimization with Laggard Data Pipelines

- Naman Agarwal, Rohan Anil, Tomer Koren, Kunal Talwar, Cyril Zhang
- abstract@[open-review](#): State-of-the-art optimization is steadily shifting towards massively parallel pipelines with extremely large batch sizes. As a consequence, CPU-bound preprocessing and disk/memory/network operations have emerged as new performance bottlenecks, as opposed to hardware-accelerated gradient computations. In this regime, a recently proposed approach is data echoing (Choi et al., 2019), which takes repeated gradient steps on the same batch while waiting for fresh data to arrive from upstream. We provide the first convergence analyses of "data-echoed" extensions of common optimization methods, showing that they exhibit provable improvements over their synchronous counterparts. Specifically, we show that in convex optimization with stochastic minibatches, data echoing affords speedups on the curvature-dominated part of the convergence rate, while maintaining the optimal statistical rate.

Self-supervised Auxiliary Learning with Meta-paths for Heterogeneous Graphs

- Dasol Hwang, Jinyoung Park, Sunyoung Kwon, KyungMin Kim, Jung-Woo Ha, Hyunwoo J. Kim
- abstract@[open-review](#): Graph neural networks have shown superior performance in a wide range of applications providing a powerful representation of graph-structured data. Recent works show that the representation can be further improved by auxiliary tasks. However, the auxiliary tasks for heterogeneous graphs, which contain rich semantic information with various types of nodes and edges, have less explored in the literature. In this paper, to learn graph neural networks on heterogeneous graphs we propose a novel self-supervised auxiliary learning method using meta paths, which are composite relations of multiple edge types. Our proposed method is learning to learn a primary task by predicting meta-paths as auxiliary tasks. This can be viewed as a type of meta-learning. The proposed method can identify an effective combination of auxiliary tasks and automatically balance them to improve the primary task. Our methods can be applied to any graph neural networks in a plug-in manner without manual labeling or additional data. The experiments demonstrate that the proposed method consistently improves the performance of link prediction and node classification on heterogeneous graphs.

GPS-Net: Graph-based Photometric Stereo Network

- Zhuokun Yao, Kun Li, Ying Fu, Haofeng Hu, Boxin Shi
- abstract@[open-review](#): Learning-based photometric stereo methods predict the surface normal either in a per-pixel or an all-pixel manner. Per-pixel methods explore the inter-image intensity variation of each pixel but ignore features from the intra-image spatial domain. All-pixel methods explore the intra-image intensity variation of each input image but pay less attention to the inter-image lighting variation. In this paper, we present a Graph-based Photometric Stereo Network, which unifies per-pixel and all-pixel processings to explore both inter-image and intra-image information. For per-pixel operation, we propose the Unstructured Feature Extraction Layer to connect an arbitrary number of input image-light pairs into graph structures, and introduce Structure-aware Graph Convolution filters to balance the input data by appropriately weighting shadows and specular highlights. For all-pixel operation, we propose the Normal Regression Network to make efficient use of the intra-image spatial information for predicting a surface normal map with rich details. Experimental results on the real-world benchmark show that our method achieves excellent performance under both sparse and dense lighting distributions.

Consistent Structural Relation Learning for Zero-Shot Segmentation

- Peike Li, Yunchao Wei, Yi Yang
- abstract@[open-review](#): Zero-shot semantic segmentation aims to recognize the semantics of pixels from unseen categories with zero training samples. Previous practice [1] proposed to train the classifiers for unseen categories using the visual features generated from semantic word embeddings. However, the generator is merely learned on the seen categories while no constraint is applied to the unseen categories, leading to poor generalization ability. In this work, we propose a Consistent Structural Relation Learning (CSRL) approach to constrain the generating of unseen visual features by exploiting the structural relations between seen and unseen categories. We observe that different categories are usually with similar relations in either semantic word embedding space or visual feature space. This observation motivates us to harness the similarity of category-level relations on the semantic word embedding space to learn a better visual feature generator. Concretely, by exploring the pair-wise and list-wise structures, we impose the relations of generated visual features to be consistent with their counterparts in the semantic word embedding space. In this way, the relations between seen and unseen categories will be transferred to implicitly constrain the generator to produce relation-consistent unseen visual features. We conduct extensive experiments on Pascal-VOC and Pascal-Context benchmarks. The proposed CSRL significantly outperforms existing state-of-the-art methods by a large margin, resulting in ~7-12% on Pascal-VOC and ~2-5% on Pascal-Context.

Model Selection in Contextual Stochastic Bandit Problems

- Aldo Pacchiano, My Phan, Yasin Abbasi Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, Csaba Szepesvari
- abstract@[open-review](#): We study bandit model selection in stochastic environments. Our approach relies on a master algorithm that selects between candidate base algorithms. We develop a master-base algorithm abstraction that can work with general classes of base algorithms and different type of adversarial master algorithms. Our methods rely on a novel and generic smoothing transformation for bandit algorithms that permits us to obtain optimal $\$O(\sqrt{T})\$$ model selection guarantees for stochastic contextual bandit problems as long as the optimal base algorithm satisfies a high probability regret guarantee. We show through a lower bound that even when one of the base algorithms has $\$O(\log T)\$$ regret, in general it is impossible to get

better than $\Omega(\sqrt{T})$ regret in model selection, even asymptotically. Using our techniques, we address model selection in a variety of problems such as misspecified linear contextual bandits [lattimore2019learning], linear bandit with unknown dimension [Foster-Krishnamurthy-Luo-2019] and reinforcement learning with unknown feature maps. Our algorithm requires the knowledge of the optimal base regret to adjust the master learning rate. We show that without such prior knowledge any master can suffer a regret larger than the optimal base regret.

[Truncated Linear Regression in High Dimensions](#)

- Constantinos Daskalakis, Dhruv Rohatgi, Emmanouil Zampetakis
- abstract@[open-review](#): As in standard linear regression, in truncated linear regression, we are given access to observations (A_i, y_i) whose dependent variable equals $y_i = A_i^\top x + \eta_i$, where x is some fixed unknown vector of interest and η_i is independent noise; except we are only given an observation if its dependent variable y_i lies in some "truncation set" $S \subset \mathbb{R}$. The goal is to recover x under some favorable conditions on the A_i 's and the noise distribution. We prove that there exists a computationally and statistically efficient method for recovering k -sparse n -dimensional vectors x from m truncated samples, which attains an optimal ℓ_2 reconstruction error of $O(\sqrt{(k \log n)/m})$. As a corollary, our guarantees imply a computationally efficient and information-theoretically optimal algorithm for compressed sensing with truncation, such as that which may arise from measurement saturation effects. Our result follows from a statistical and computational analysis of the Stochastic Gradient Descent (SGD) algorithm for solving a natural adaption of the LASSO optimization problem that accommodates truncation. This generalizes the works of both: (1) [Daskalakis et al. 2018], where no regularization is needed due to the low dimensionality of the data, and (2) [Wainright 2009], where the objective function is simple due to the absence of truncation. In order to deal with both truncation and high-dimensionality at the same time, we develop new techniques that not only generalize the existing ones but we believe are of independent interest.

[Incorporating Pragmatic Reasoning Communication into Emergent Language](#)

- Yipeng Kang, Tonghan Wang, Gerard de Melo
- abstract@[open-review](#): Emergentism and pragmatics are two research fields that study the dynamics of linguistic communication along quite different timescales and intelligence levels. From the perspective of multi-agent reinforcement learning, they correspond to stochastic games with reinforcement training and stage games with opponent awareness, respectively. Given that their combination has been explored in linguistics, in this work, we combine computational models of short-term mutual reasoning-based pragmatics with long-term language emergentism. We explore this for agent communication in two settings, referential games and Starcraft II, assessing the relative merits of different kinds of mutual reasoning pragmatics models both empirically and theoretically. Our results shed light on their importance for making inroads towards getting more natural, accurate, robust, fine-grained, and succinct utterances.

[Deep Subspace Clustering with Data Augmentation](#)

- Mahdi Abavisani, Alireza Naghizadeh, Dimitris Metaxas, Vishal Patel
- abstract@[open-review](#): The idea behind data augmentation techniques is based on the fact that slight changes in the percept do not change the brain cognition. In classification, neural networks use this fact by applying transformations to the inputs to learn to predict the same label. However, in deep subspace clustering (DSC), the ground-truth labels are not available, and as a result, one cannot easily use data augmentation techniques. We propose a technique to exploit the benefits of data augmentation in DSC algorithms. We learn representations that have consistent subspaces for slightly transformed inputs. In particular, we introduce a temporal ensembling component to the objective function of DSC algorithms to enable the DSC networks to maintain consistent subspaces for random transformations in the input data. In addition, we provide a simple yet effective unsupervised procedure to find efficient data augmentation policies. An augmentation policy is defined as an image processing transformation with a certain magnitude and probability of being applied to each image in each epoch. We search through the policies in a search space of the most common augmentation policies to find the best policy such that the DSC network yields the highest mean Silhouette coefficient in its clustering results on a target dataset. Our method achieves state-of-the-art performance on four standard subspace clustering datasets.

[An Empirical Process Approach to the Union Bound: Practical Algorithms for Combinatorial and Linear Bandits](#)

- Julian Katz-Samuels, Lalit Jain, zohar karnin, Kevin G. Jamieson
- abstract@[open-review](#): This paper proposes near-optimal algorithms for the pure-exploration linear bandit problem in the fixed confidence and fixed budget settings. Leveraging ideas from the theory of suprema of empirical processes, we provide an algorithm whose sample complexity scales with the geometry of the instance and avoids an explicit union bound over the number of arms. Unlike previous approaches which sample based on minimizing a worst-case variance (e.g. G-optimal design), we define an experimental design objective based on the Gaussian-width of the underlying arm set. We provide a novel lower bound in terms of this objective that highlights its fundamental role in the sample complexity. The sample complexity of our fixed confidence algorithm matches this lower bound, and in addition is computationally efficient for combinatorial classes, e.g. shortest-path, matchings and matroids, where the arm sets can be exponentially large in the dimension. Finally, we propose the first algorithm for linear bandits in the the fixed budget setting. Its guarantee matches our lower bound up to logarithmic factors.

[Can Graph Neural Networks Count Substructures?](#)

- Zhengdao Chen, Lei Chen, Soledad Villar, Joan Bruna
- abstract@[open-review](#): The ability to detect and count certain substructures in graphs is important for solving many tasks on graph-structured data, especially in the contexts of computational chemistry and biology as well as social network analysis. Inspired by this, we propose to study the expressive power of graph neural networks (GNNs) via their ability to count attributed graph substructures, extending recent works that examine their power in graph isomorphism testing and function approximation. We distinguish between two types of substructure counting: induced-subgraph-count and subgraph-count, and establish both positive and negative answers for popular GNN architectures. Specifically, we prove that Message Passing Neural Networks (MPNNs), 2-Weisfeiler-Lehman (2-WL) and 2-Invariant Graph Networks (2-IGNs) cannot perform induced-subgraph-count of substructures consisting of 3 or more nodes, while they can perform subgraph-count of star-shaped substructures. As an intermediary step, we prove that 2-WL and 2-IGNs are equivalent in distinguishing non-isomorphic graphs, partly answering an open problem raised in Maron et al. (2019). We also prove positive results for k-WL and k-IGNs as well as negative results for k-WL with a finite number of iterations. We then conduct experiments that support the theoretical results for MPNNs and 2-IGNs. Moreover, motivated by substructure counting and inspired by Murphy et al. (2019), we propose the Local Relational Pooling model and demonstrate that it is not only effective for substructure counting but also able to achieve competitive performance on molecular prediction tasks.

[A Bayesian Perspective on Training Speed and Model Selection](#)

- Clare Lyle, Lisa Schut, Robin Ru, Yarin Gal, Mark van der Wilk
- abstract@[open-review](#): We take a Bayesian perspective to illustrate a connection between training speed and the marginal likelihood in linear models. This provides two major insights: first, that a measure of a model's training speed can be used to estimate its marginal likelihood. Second, that this measure, under certain conditions, predicts the relative weighting of models in linear model combinations trained to minimize a regression loss. We verify our results in model selection tasks for linear models and for the infinite-width limit of deep neural networks. We further provide encouraging empirical evidence that the intuition developed in these settings also holds for deep neural networks trained with stochastic gradient descent. Our results suggest a

promising new direction towards explaining why neural networks trained with stochastic gradient descent are biased towards functions that generalize well.

[On the Modularity of Hypernetworks](#)

- Tomer Galanti, Lior Wolf
- abstract@[open-review](#): In the context of learning to map an input I to a function $h_I: \mathcal{X} \rightarrow \mathbb{R}$, two alternative methods are compared: (i) an embedding-based method, which learns a fixed function in which I is encoded as a conditioning signal $e(I)$ and the learned function takes the form $h_I(x) = q(x, e(I))$, and (ii) hypernetworks, in which the weights θ_I of the function $h_I(x) = g(x; \theta_I)$ are given by a hypernetwork f as $\theta_I = f(I)$. In this paper, we define the property of modularity as the ability to effectively learn a different function for each input instance I . For this purpose, we adopt an expressivity perspective of this property and extend the theory of [devore](#) and provide a lower bound on the complexity (number of trainable parameters) of neural networks as function approximators, by eliminating the requirements for the approximation method to be robust. Our results are then used to compare the complexities of q and g , showing that under certain conditions and when letting the functions e and f be as large as we wish, g can be smaller than q by orders of magnitude. This sheds light on the modularity of hypernetworks in comparison with the embedding-based method. Besides, we show that for a structured target function, the overall number of trainable parameters in a hypernetwork is smaller by orders of magnitude than the number of trainable parameters of a standard neural network and an embedding method.

[Doubly Robust Off-Policy Value and Gradient Estimation for Deterministic Policies](#)

- Nathan Kallus, Masatoshi Uehara
- abstract@[open-review](#): Offline reinforcement learning, wherein one uses off-policy data logged by a fixed behavior policy to evaluate and learn new policies, is crucial in applications where experimentation is limited such as medicine. We study the estimation of policy value and gradient of a deterministic policy from off-policy data when actions are continuous. Targeting deterministic policies, for which action is a deterministic function of state, is crucial since optimal policies are always deterministic (up to ties). In this setting, standard importance sampling and doubly robust estimators for policy value and gradient fail because the density ratio does not exist. To circumvent this issue, we propose several new doubly robust estimators based on different kernelization approaches. We analyze the asymptotic mean-squared error of each of these under mild rate conditions for nuisance estimators. Specifically, we demonstrate how to obtain a rate that is independent of the horizon length.

[Provably Efficient Neural GTD for Off-Policy Learning](#)

- Hoi-To Wai, Zhuoran Yang, Zhaoran Wang, Mingyi Hong
- abstract@[open-review](#): This paper studies a gradient temporal difference (GTD) algorithm using neural network (NN) function approximators to minimize the mean squared Bellman error (MSBE). For off-policy learning, we show that the minimum MSBE problem can be recast into a min-max optimization involving a pair of over-parameterized primal-dual NNs. The resultant formulation can then be tackled using a neural GTD algorithm. We analyze the convergence of the proposed algorithm with a 2-layer ReLU NN architecture using m neurons and prove that it computes an approximate optimal solution to the minimum MSBE problem as $m \rightarrow \infty$.

[Learning Discrete Energy-based Models via Auxiliary-variable Local Exploration](#)

- Hanjun Dai, Rishabh Singh, Bo Dai, Charles Sutton, Dale Schuurmans
- abstract@[open-review](#): Discrete structures play an important role in applications like program language modeling and software engineering. Current approaches to predicting complex structures typically consider autoregressive models for their tractability, with some sacrifice in flexibility. Energy-based models (EBMs) on the other hand offer a more flexible and thus more powerful approach to modeling such distributions, but require partition function estimation. In this paper we propose `modelshort`, a new algorithm for learning conditional and unconditional EBMs for discrete structured data, where parameter gradients are estimated using a learned sampler that mimics local search. We show that the energy function and sampler can be trained efficiently via a new variational form of power iteration, achieving a better trade-off between flexibility and tractability. Experimentally, we show that learning local search leads to significant improvements in challenging application domains. Most notably, we present an energy model guided fuzzer for software testing that achieves comparable performance to well engineered fuzzing engines like libfuzzer.

[Stable and expressive recurrent vision models](#)

- Drew Linsley, Alekh Karkada Ashok, Lakshmi Narasimhan Govindarajan, Rex Liu, Thomas Serre
- abstract@[open-review](#): Primate vision depends on recurrent processing for reliable perception. A growing body of literature also suggests that recurrent connections improve the learning efficiency and generalization of vision models on classic computer vision challenges. Why then, are current large-scale challenges dominated by feedforward networks? We posit that the effectiveness of recurrent vision models is bottlenecked by the standard algorithm used for training them, "back-propagation through time" (BPTT), which has $O(N)$ memory-complexity for training an N step model. Thus, recurrent vision model design is bounded by memory constraints, forcing a choice between rivaling the enormous capacity of leading feedforward models or trying to compensate for this deficit through granular and complex dynamics. Here, we develop a new learning algorithm, "contractor recurrent back-propagation" (C-RBP), which alleviates these issues by achieving constant $O(1)$ memory-complexity with steps of recurrent processing. We demonstrate that recurrent vision models trained with C-RBP can detect long-range spatial dependencies in a synthetic contour tracing task that BPTT-trained models cannot. We further show that recurrent vision models trained with C-RBP to solve the large-scale Panoptic Segmentation MS-COCO challenge outperform the leading feedforward approach, with fewer free parameters. C-RBP is a general-purpose learning algorithm for any application that can benefit from expansive recurrent dynamics. Code and data are available at <https://github.com/c-rbp>.

[Entropic Optimal Transport between Unbalanced Gaussian Measures has a Closed Form](#)

- Hicham Janati, Boris Muzellec, Gabriel PeyrÃ©, Marco Cuturi
- abstract@[open-review](#): Although optimal transport (OT) problems admit closed form solutions in a very few notable cases, e.g. in 1D or between Gaussians, these closed forms have proved extremely fecund for practitioners to define tools inspired from the OT geometry. On the other hand, the numerical resolution of OT problems using entropic regularization has given rise to many applications, but because there are no known closed-form solutions for entropic regularized OT problems, these approaches are mostly algorithmic, not informed by elegant closed forms. In this paper, we propose to fill the void at the intersection between these two schools of thought in OT by proving that the entropy-regularized optimal transport problem between two Gaussian measures admits a closed form. Contrary to the unregularized case, for which the explicit form is given by the Wasserstein-Bures distance, the closed form we obtain is differentiable everywhere, even for Gaussians with degenerate covariance matrices. We obtain this closed form solution by solving the fixed-point equation behind Sinkhorn's algorithm, the default method for computing entropic regularized OT. Remarkably, this approach extends to the generalized unbalanced case --- where Gaussian measures are scaled by positive constants. This extension leads to a closed form expression for unbalanced Gaussians as well, and highlights the mass transportation / destruction trade-off seen in unbalanced optimal transport. Moreover, in both settings, we show that the optimal transportation plans are (scaled) Gaussians and provide analytical formulas of their parameters. These formulas constitute the first non-trivial closed forms for entropy-regularized optimal transport, thus providing a ground truth for the analysis of entropic OT and Sinkhorn's algorithm.

[BRP-NAS: Prediction-based NAS using GCNs](#)

- Lukasz Dudziak, Thomas Chau, Mohamed Abdelfattah, Royson Lee, Hyeji Kim, Nicholas Lane
- abstract@[open-review](#): Neural architecture search (NAS) enables researchers to automatically explore broad design spaces in order to improve efficiency of neural networks. This efficiency is especially important in the case of on-device deployment, where improvements in accuracy should be balanced out with computational demands of a model. In practice, performance metrics of model are computationally expensive to obtain. Previous work uses a proxy (e.g., number of operations) or a layer-wise measurement of neural network layers to estimate end-to-end hardware performance but the imprecise prediction diminishes the quality of NAS. To address this problem, we propose BRP-NAS, an efficient hardware-aware NAS enabled by an accurate performance predictor-based on graph convolutional network (GCN). What is more, we investigate prediction quality on different metrics and show that sample efficiency of the predictor-based NAS can be improved by considering binary relations of models and an iterative data selection strategy. We show that our proposed method outperforms all prior methods on NAS-Bench-101, NAS-Bench-201 and DARTS. Finally, to raise awareness of the fact that accurate latency estimation is not a trivial task, we release LatBench -- a latency dataset of NAS-Bench-201 models running on a broad range of devices.

[Deep Shells: Unsupervised Shape Correspondence with Optimal Transport](#)

- Marvin Eisenberger, Aysim Toker, Laura Leal-Taix, Daniel Cremers
- abstract@[open-review](#): We propose a novel unsupervised learning approach to 3D shape correspondence that builds a multiscale matching pipeline into a deep neural network. This approach is based on smooth shells, the current state-of-the-art axiomatic correspondence method, which requires an a priori stochastic search over the space of initial poses. Our goal is to replace this costly preprocessing step by directly learning good initializations from the input surfaces. To that end, we systematically derive a fully differentiable, hierarchical matching pipeline from entropy regularized optimal transport. This allows us to combine it with a local feature extractor based on smooth, truncated spectral convolution filters. Finally, we show that the proposed unsupervised method significantly improves over the state-of-the-art on multiple datasets, even in comparison to the most recent supervised methods. Moreover, we demonstrate compelling generalization results by applying our learned filters to examples that significantly deviate from the training set.

[ISTA-NAS: Efficient and Consistent Neural Architecture Search by Sparse Coding](#)

- Yibo Yang, Hongyang Li, Shan You, Fei Wang, Chen Qian, Zhouchen Lin
- abstract@[open-review](#): Neural architecture search (NAS) aims to produce the optimal sparse solution from a high-dimensional space spanned by all candidate connections. Current gradient-based NAS methods commonly ignore the constraint of sparsity in the search phase, but project the optimized solution onto a sparse one by post-processing. As a result, the dense super-net for search is inefficient to train and has a gap with the projected architecture for evaluation. In this paper, we formulate neural architecture search as a sparse coding problem. We perform the differentiable search on a compressed lower-dimensional space that has the same validation loss as the original sparse solution space, and recover an architecture by solving the sparse coding problem. The differentiable search and architecture recovery are optimized in an alternate manner. By doing so, our network for search at each update satisfies the sparsity constraint and is efficient to train. In order to also eliminate the depth and width gap between the network in search and the target-net in evaluation, we further propose a method to search and evaluate in one stage under the target-net settings. When training finishes, architecture variables are absorbed into network weights. Thus we get the searched architecture and optimized parameters in a single run. In experiments, our two-stage method on CIFAR-10 requires only 0.05 GPU-day for search. Our one-stage method produces state-of-the-art performances on both CIFAR-10 and ImageNet at the cost of only evaluation time.

[Rel3D: A Minimally Contrastive Benchmark for Grounding Spatial Relations in 3D](#)

- Ankit Goyal, Kaiyu Yang, Dawei Yang, Jia Deng
- abstract@[open-review](#): Understanding spatial relations (e.g., laptop on table) in visual input is important for both humans and robots. Existing datasets are insufficient as they lack large-scale, high-quality 3D ground truth information, which is critical for learning spatial relations. In this paper, we fill this gap by constructing Rel3D: the first large-scale, human-annotated dataset for grounding spatial relations in 3D. Rel3D enables quantifying the effectiveness of 3D information in predicting spatial relations on large-scale human data. Moreover, we propose minimally contrastive data collection---a novel crowdsourcing method for reducing dataset bias. The 3D scenes in our dataset come in minimally contrastive pairs: two scenes in a pair are almost identical, but a spatial relation holds in one and fails in the other. We empirically validate that minimally contrastive examples can diagnose issues with current relation detection models as well as lead to sample-efficient training. Code and data are available at <https://github.com/princeton-vl/Rel3D>.

[Regularizing Black-box Models for Improved Interpretability](#)

- Gregory Plumb, Maruan Al-Shedivat, Ángel Alexander Cabrera, Adam Perer, Eric Xing, Ameet Talwalkar
- abstract@[open-review](#): Most of the work on interpretable machine learning has focused on designing either inherently interpretable models, which typically trade-off accuracy for interpretability, or post-hoc explanation systems, whose explanation quality can be unpredictable. Our method, ExpO, is a hybridization of these approaches that regularizes a model for explanation quality at training time. Importantly, these regularizers are differentiable, model agnostic, and require no domain knowledge to define. We demonstrate that post-hoc explanations for ExpO-regularized models have better explanation quality, as measured by the common fidelity and stability metrics. We verify that improving these metrics leads to significantly more useful explanations with a user study on a realistic task.

[Trust the Model When It Is Confident: Masked Model-based Actor-Critic](#)

- Feiyang Pan, Jia He, Dandan Tu, Qing He
- abstract@[open-review](#): It is a popular belief that model-based Reinforcement Learning (RL) is more sample efficient than model-free RL, but in practice, it is not always true due to overweighed model errors. In complex and noisy settings, model-based RL tends to have trouble using the model if it does not know when to trust the model. In this work, we find that better model usage can make a huge difference. We show theoretically that if the use of model-generated data is restricted to state-action pairs where the model error is small, the performance gap between model and real rollouts can be reduced. It motivates us to use model rollouts only when the model is confident about its predictions. We propose Masked Model-based Actor-Critic (M2AC), a novel policy optimization algorithm that maximizes a model-based lower-bound of the true value function. M2AC implements a masking mechanism based on the model's uncertainty estimation to decide whether the model should be used or not. Consequently, the new algorithm tends to give robust policy improvements. Experiments on continuous control benchmarks demonstrate that M2AC has strong performance even when using long model rollouts in very noisy environments, and significantly outperforms previous state-of-the-art methods.

[Semi-Supervised Neural Architecture Search](#)

- Renqian Luo, Xu Tan, Rui Wang, Tao Qin, Enhong Chen, Tie-Yan Liu
- abstract@[open-review](#): Neural architecture search (NAS) relies on a good controller to generate better architectures or predict the accuracy of given architectures. However, training the controller requires both abundant and high-quality pairs of architectures and their accuracy, while it is costly to evaluate an architecture and obtain its accuracy. In this paper, we propose SemiNAS, a semi-supervised NAS approach that leverages numerous unlabeled architectures (without evaluation and thus nearly no cost). Specifically, SemiNAS 1) trains an initial accuracy predictor with a small set of architecture-accuracy data pairs; 2) uses the trained accuracy predictor to predict the accuracy of large amount of architectures (without evaluation); and 3) adds the

generated data pairs to the original data to further improve the predictor. The trained accuracy predictor can be applied to various NAS algorithms by predicting the accuracy of candidate architectures for them. SemiNAS has two advantages: 1) It reduces the computational cost under the same accuracy guarantee. On NASBench-101 benchmark dataset, it achieves comparable accuracy with gradient-based method while using only 1/7 architecture-accuracy pairs. 2) It achieves higher accuracy under the same computational cost. It achieves 94.02% test accuracy on NASBench-101, outperforming all the baselines when using the same number of architectures. On ImageNet, it achieves 23.5% top-1 error rate (under 600M FLOPS constraint) using 4 GPU-days for search. We further apply it to LJSpeech text to speech task and it achieves 97% intelligibility rate in the low-resource setting and 15% test error rate in the robustness setting, with 9%, 7% improvements over the baseline respectively.

[Consistency Regularization for Certified Robustness of Smoothed Classifiers](#)

- Jongheon Jeong, Jinwoo Shin
- abstract@[open-review](#): A recent technique of randomized smoothing has shown that the worst-case (adversarial) ℓ_2 -robustness can be transformed into the average-case Gaussian-robustness by "smoothing" a classifier, i.e., by considering the averaged prediction over Gaussian noise. In this paradigm, one should rethink the notion of adversarial robustness in terms of generalization ability of a classifier under noisy observations. We found that the trade-off between accuracy and certified robustness of smoothed classifiers can be greatly controlled by simply regularizing the prediction consistency over noise. This relationship allows us to design a robust training objective without approximating a non-existing smoothed classifier, e.g., via soft smoothing. Our experiments under various deep neural network architectures and datasets show that the "certified" ℓ_2 -robustness can be dramatically improved with the proposed regularization, even achieving better or comparable results to the state-of-the-art approaches with significantly less training costs and hyperparameters.

[Robust Multi-Agent Reinforcement Learning with Model Uncertainty](#)

- Kaiqing Zhang, TAO SUN, Yunzhe Tao, Sahika Genc, Sunil Mallya, Tamer Basar
- abstract@[open-review](#): In this work, we study the problem of multi-agent reinforcement learning (MARL) with model uncertainty, which is referred to as robust MARL. This is naturally motivated by some multi-agent applications where each agent may not have perfectly accurate knowledge of the model, e.g., all the reward functions of other agents. Little a priori work on MARL has accounted for such uncertainties, neither in problem formulation nor in algorithm design. In contrast, we model the problem as a robust Markov game, where the goal of all agents is to find policies such that no agent has the incentive to deviate, i.e., reach some equilibrium point, which is also robust to the possible uncertainty of the MARL model. We first introduce the solution concept of robust Nash equilibrium in our setting, and develop a Q-learning algorithm to find such equilibrium policies, with convergence guarantees under certain conditions. In order to handle possibly enormous state-action spaces in practice, we then derive the policy gradients for robust MARL, and develop an actor-critic algorithm with function approximation. Our experiments demonstrate that the proposed algorithm outperforms several baseline MARL methods that do not account for the model uncertainty, in several standard but uncertain cooperative and competitive MARL environments.

[SIRI: Spatial Relation Induced Network For Spatial Description Resolution](#)

- peiyao wang, Weixin Luo, Yanyu Xu, Haojie Li, Shugong Xu, Jianyu Yang, Shenghua Gao
- abstract@[open-review](#): Spatial Description Resolution, as a language-guided localization task, is proposed for target location in a panoramic street view, given corresponding language descriptions. Explicitly characterizing an object-level relationship while distilling spatial relationships are currently absent but crucial to this task. Mimicking humans, who sequentially traverse spatial relationship words and objects with a first-person view to locate their target, we propose a novel spatial relationship induced (SIRI) network. Specifically, visual features are firstly correlated at an implicit object-level in a projected latent space; then they are distilled by each spatial relationship word, resulting in each differently activated feature representing each spatial relationship. Further, we introduce global position priors to fix the absence of positional information, which may result in global positional reasoning ambiguities. Both the linguistic and visual features are concatenated to finalize the target localization. Experimental results on the Touchdown show that our method is around 24% better than the state-of-the-art method in terms of accuracy, measured by an 80-pixel radius. Our method also generalizes well on our proposed extended dataset collected using the same settings as Touchdown. The code for this project is publicly available at <https://github.com/wong-puiyiu/siri-sdr>.

[Adaptive Shrinkage Estimation for Streaming Graphs](#)

- Nesreen Ahmed, Nick Duffield
- abstract@[open-review](#): Networks are a natural representation of complex systems across the sciences, and higher-order dependencies are central to the understanding and modeling of these systems. However, in many practical applications such as online social networks, networks are massive, dynamic, and naturally streaming, where pairwise interactions among vertices become available one at a time in some arbitrary order. The massive size and streaming nature of these networks allow only partial observation, since it is infeasible to analyze the entire network. Under such scenarios, it is challenging to study the higher-order structural and connectivity patterns of streaming networks. In this work, we consider the fundamental problem of estimating the higher-order dependencies using adaptive sampling. We propose a novel adaptive, single-pass sampling framework and unbiased estimators for higher-order network analysis of large streaming networks. Our algorithms exploit adaptive techniques to identify edges that are highly informative for efficiently estimating the higher-order structure of streaming networks from small sample data. We also introduce a novel James-Stein shrinkage estimator to reduce the estimation error. Our approach is fully analytic, computationally efficient, and can be incrementally updated in a streaming setting. Numerical experiments on large networks show that our approach is superior to baseline methods.

[Make One-Shot Video Object Segmentation Efficient Again](#)

- Tim Meinhardt, Laura Leal-Taixé
- abstract@[open-review](#): Video object segmentation (VOS) describes the task of segmenting a set of objects in each frame of a video. In the semi-supervised setting, the first mask of each object is provided at test time. Following the one-shot principle, fine-tuning VOS methods train a segmentation model separately on each given object mask. However, recently the VOS community has deemed such a test time optimization and its impact on the test runtime as unfeasible. To mitigate the inefficiencies of previous fine-tuning approaches, we present efficient One-Shot Video Object Segmentation (e-OSVOS). In contrast to most VOS approaches, e-OSVOS decouples the object detection task and predicts only local segmentation masks by applying a modified version of Mask R-CNN. The one-shot test runtime and performance are optimized without a laborious and handcrafted hyperparameter search. To this end, we meta learn the model initialization and learning rates for the test time optimization. To achieve an optimal learning behavior, we predict individual learning rates at a neuron level. % a pair of learning rates for the weights tensor and scalar bias of each neuron. Furthermore, we apply an online adaptation to address the common performance degradation throughout a sequence by continuously fine-tuning the model on previous mask predictions supported by a frame-to-frame bounding box propagation. % through changing online appearance -> online adaptation for free. bounding box propagation. e-OSVOS provides state-of-the-art results on DAVIS 2016, DAVIS 2017 and YouTube-VOS for one-shot fine-tuning methods while reducing the test runtime substantially.

[Depth Uncertainty in Neural Networks](#)

- Javier Antoran, James Allingham, José Miguel Hernández-Lobato

- abstract@[open-review](#): Existing methods for estimating uncertainty in deep learning tend to require multiple forward passes, making them unsuitable for applications where computational resources are limited. To solve this, we perform probabilistic reasoning over the depth of neural networks. Different depths correspond to subnetworks which share weights and whose predictions are combined via marginalisation, yielding model uncertainty. By exploiting the sequential structure of feed-forward networks, we are able to both evaluate our training objective and make predictions with a single forward pass. We validate our approach on real-world regression and image classification tasks. Our approach provides uncertainty calibration, robustness to dataset shift, and accuracies competitive with more computationally expensive baselines.

[Non-Euclidean Universal Approximation](#)

- Anastasis Kratsios, Ievgen Bilokopytov
- abstract@[open-review](#): Modifications to a neural network's input and output layers are often required to accommodate the specificities of most practical learning tasks. However, the impact of such changes on architecture's approximation capabilities is largely not understood. We present general conditions describing feature and readout maps that preserve an architecture's ability to approximate any continuous functions uniformly on compacts. As an application, we show that if an architecture is capable of universal approximation, then modifying its final layer to produce binary values creates a new architecture capable of deterministically approximating any classifier. In particular, we obtain guarantees for deep CNNs, deep ffNN, and universal Gaussian processes. Our results also have consequences within the scope of geometric deep learning. Specifically, when the input and output spaces are Hadamard manifolds, we obtain geometrically meaningful feature and readout maps satisfying our criteria. Consequently, commonly used non-Euclidean regression models between spaces of symmetric positive definite matrices are extended to universal DNNs. The same result allows us to show that the hyperbolic feed-forward networks, used for hierarchical learning, are universal. Our result is also used to show that the common practice of randomizing all but the last two layers of a DNN produces a universal family of functions with probability one. We also provide conditions on a DNN's first (resp. last) few layer's connections and activation function which guarantee that these layer's can have a width equal to the input (resp. output) space's dimension while not negatively effecting the architecture's approximation capabilities.

[Constraining Variational Inference with Geometric Jensen-Shannon Divergence](#)

- Jacob Deasy, Nikola Simidjievski, Pietro Līo³
- abstract@[open-review](#): We examine the problem of controlling divergences for latent space regularisation in variational autoencoders. Specifically, when aiming to reconstruct example $\mathbf{x} \in \mathbb{R}^m$ via latent space $\mathbf{z} \in \mathbb{R}^n$ ($n \leq m$), while balancing this against the need for generalisable latent representations. We present a regularisation mechanism based on the $\text{skew-geometric Jensen-Shannon divergence}$ $\left(\text{JS}(\mathbf{G}|\mathbf{G}) \right)$. We find a variation in $\text{JS}(\mathbf{G}|\mathbf{G})$, motivated by limiting cases, which leads to an intuitive interpolation between forward and reverse KL in the space of both distributions and divergences. We motivate its potential benefits for VAEs through low-dimensional examples, before presenting quantitative and qualitative results. Our experiments demonstrate that skewing our variant of $\text{JS}(\mathbf{G}|\mathbf{G})$, in the context of $\text{JS}(\mathbf{G}|\mathbf{G})$ -VAEs, leads to better reconstruction and generation when compared to several baseline VAEs. Our approach is entirely unsupervised and utilises only one hyperparameter which can be easily interpreted in latent space.

[Gibbs Sampling with People](#)

- Peter Harrison, Raja Marjieh, Federico Adolfi, Pol van Rijn, Manuel Anglada-Tort, Ofer Tchernichovski, Pauline Larrouy-Maestri, Nori Jacoby
- abstract@[open-review](#): A core problem in cognitive science and machine learning is to understand how humans derive semantic representations from perceptual objects, such as color from an apple, pleasantness from a musical chord, or seriousness from a face. Markov Chain Monte Carlo with People (MCMCP) is a prominent method for studying such representations, in which participants are presented with binary choice trials constructed such that the decisions follow a Markov Chain Monte Carlo acceptance rule. However, while MCMCP has strong asymptotic properties, its binary choice paradigm generates relatively little information per trial, and its local proposal function makes it slow to explore the parameter space and find the modes of the distribution. Here we therefore generalize MCMCP to a continuous-sampling paradigm, where in each iteration the participant uses a slider to continuously manipulate a single stimulus dimension to optimize a given criterion such as "pleasantness". We formulate both methods from a utility-theory perspective, and show that the new method can be interpreted as "Gibbs Sampling with People" (GSP). Further, we introduce an aggregation parameter to the transition step, and show that this parameter can be manipulated to flexibly shift between Gibbs sampling and deterministic optimization. In an initial study, we show GSP clearly outperforming MCMCP; we then show that GSP provides novel and interpretable results in three other domains, namely musical chords, vocal emotions, and faces. We validate these results through large-scale perceptual rating experiments. The final experiments use GSP to navigate the latent space of a state-of-the-art image synthesis network (StyleGAN), a promising approach for applying GSP to high-dimensional perceptual spaces. We conclude by discussing future cognitive applications and ethical implications.

[HM-ANN: Efficient Billion-Point Nearest Neighbor Search on Heterogeneous Memory](#)

- Jie Ren, Minjia Zhang, Dong Li
- abstract@[open-review](#): The state-of-the-art approximate nearest neighbor search (ANNS) algorithms face a fundamental tradeoff between query latency and accuracy, because of small main memory capacity: To store indices in main memory for short query latency, the ANNS algorithms have to limit dataset size or use a quantization scheme which hurts search accuracy. The emergence of heterogeneous memory (HM) brings a solution to significantly increase memory capacity and break the above tradeoff: Using HM, billions of data points can be placed in the main memory on a single machine without using any data compression. However, HM consists of both fast (but small) memory and slow (but large) memory, and using HM inappropriately slows down query significantly. In this work, we present a novel graph-based similarity search algorithm called HM-ANN, which takes both memory and data heterogeneity into consideration and enables billion-scale similarity search on a single node without using compression. On two billion-sized datasets BIGANN and DEEP1B, HM-ANN outperforms state-of-the-art compression-based solutions such as L&C and IMI+OPQ in recall-vs-latency by a large margin, obtaining 46% higher recall under the same search latency. We also extend existing graph-based methods such as HNSW and NSG with two strong baseline implementations on HM. At billion-point scale, HM-ANN is 2X and 5.8X faster than our HNSW and NSG baselines respectively to reach the same accuracy.

[FrugalML: How to use ML Prediction APIs more accurately and cheaply](#)

- Lingjiao Chen, Matei Zaharia, James Y. Zou
- abstract@[open-review](#): Offering prediction APIs for fee is a fast growing industry and is an important aspect of machine learning as a service. While many such services are available, the heterogeneity in their price and performance makes it challenging for users to decide which API or combination of APIs to use for their own data and budget. We take a first step towards addressing this challenge by proposing FrugalML, a principled framework that jointly learns the strength and weakness of each API on different data, and performs an efficient optimization to automatically identify the best sequential strategy to adaptively use the available APIs within a budget constraint. Our theoretical analysis shows that natural sparsity in the formulation can be leveraged to make FrugalML efficient. We conduct systematic experiments using ML APIs from Google, Microsoft, Amazon, IBM, Baidu and other providers for tasks including facial emotion recognition, sentiment analysis and speech recognition. Across various tasks, FrugalML can achieve up to 90% cost reduction while matching the accuracy of the best single API, or up to 5% better accuracy while matching the best API's cost.

[Sharp Representation Theorems for ReLU Networks with Precise Dependence on Depth](#)

- Guy Bresler, Dheeraj Nagaraj
- abstract@[open-review](#): We prove dimension free representation results for neural networks with D ReLU layers under square loss for a class of functions G_D defined in the paper. These results capture the precise benefits of depth in the following sense:
 1. The rates for representing the class of functions G_D via D ReLU layers is sharp up to constants, as shown by matching lower bounds. 2. G_D is a proper subset of G_{D+1} and as D grows the class of functions G_D grow to contain less smooth functions.
 2. If $D^{\prime} < D$, then the approximation rate achieved by depth D^{\prime} networks is strictly worse than that achieved by depth D networks for the class G_D .

This constitutes a fine-grained characterization of the representation power of feedforward networks of arbitrary depth D and number of neurons N, in contrast to existing representation results which either require D growing quickly with N or assume that the function being represented is highly smooth. In the latter case similar rates can be obtained with a single nonlinear layer. Our results confirm the prevailing hypothesis that deeper networks are better at representing less smooth functions, and indeed, the main technical novelty is to fully exploit the fact that deep networks can produce highly oscillatory functions with few activation functions.

[Shared Experience Actor-Critic for Multi-Agent Reinforcement Learning](#)

- Filippos Christianos, Lukas Schäfer, Stefano Albrecht
- abstract@[open-review](#): Exploration in multi-agent reinforcement learning is a challenging problem, especially in environments with sparse rewards. We propose a general method for efficient exploration by sharing experience amongst agents. Our proposed algorithm, called shared Experience Actor-Critic(SEAC), applies experience sharing in an actor-critic framework by combining the gradients of different agents. We evaluate SEAC in a collection of sparse-reward multi-agent environments and find that it consistently outperforms several baselines and state-of-the-art algorithms by learning in fewer steps and converging to higher returns. In some harder environments, experience sharing makes the difference between learning to solve the task and not learning at all.

[Monotone operator equilibrium networks](#)

- Ezra Winston, J. Zico Kolter
- abstract@[open-review](#): Implicit-depth models such as Deep Equilibrium Networks have recently been shown to match or exceed the performance of traditional deep networks while being much more memory efficient. However, these models suffer from unstable convergence to a solution and lack guarantees that a solution exists. On the other hand, Neural ODEs, another class of implicit-depth models, do guarantee existence of a unique solution but perform poorly compared with traditional networks. In this paper, we develop a new class of implicit-depth model based on the theory of monotone operators, the Monotone Operator Equilibrium Network (monDEQ). We show the close connection between finding the equilibrium point of an implicit network and solving a form of monotone operator splitting problem, which admits efficient solvers with guaranteed, stable convergence. We then develop a parameterization of the network which ensures that all operators remain monotone, which guarantees the existence of a unique equilibrium point. Finally, we show how to instantiate several versions of these models, and implement the resulting iterative solvers, for structured linear operators such as multi-scale convolutions. The resulting models vastly outperform the Neural ODE-based models while also being more computationally efficient. Code is available at <http://github.com/locuslab/monotoneopnet>.

[When and How to Lift the Lockdown? Global COVID-19 Scenario Analysis and Policy Assessment using Compartmental Gaussian Processes](#)

- Zhaozhi Qian, Ahmed M. Alaa, Mihaela van der Schaar
- abstract@[open-review](#): The coronavirus disease 2019 (COVID-19) global pandemic has led many countries to impose unprecedented lockdown measures in order to slow down the outbreak. Questions on whether governments have acted promptly enough, and whether lockdown measures can be lifted soon have since been central in public discourse. Data-driven models that predict COVID-19 fatalities under different lockdown policy scenarios are essential for addressing these questions, and for informing governments on future policy directions. To this end, this paper develops a Bayesian model for predicting the effects of COVID-19 containment policies in a global context — we treat each country as a distinct data point, and exploit variations of policies across countries to learn country-specific policy effects. Our model utilizes a two-layer Gaussian process (GP) prior — the lower layer uses a compartmental SEIR (Susceptible, Exposed, Infected, Recovered) model as a prior mean function with country-and-policy-specific parameters that capture fatality curves under different counterfactual policies within each country, whereas the upper layer is shared across all countries, and learns lower-layer SEIR parameters as a function of country features and policy indicators. Our model combines the solid mechanistic foundations of SEIR models (Bayesian priors) with the flexible data-driven modeling and gradient-based optimization routines of machine learning (Bayesian posteriors) — i.e., the entire model is trained end-to-end via stochastic variational inference. We compare the projections of our model with other models listed by the Center for Disease Control (CDC), and provide scenario analyses for various lockdown and reopening strategies highlighting their impact on COVID-19 fatalities.

[Unsupervised Learning of Lagrangian Dynamics from Images for Prediction and Control](#)

- Yaofeng Desmond Zhong, Naomi Leonard
- abstract@[open-review](#): Recent approaches for modelling dynamics of physical systems with neural networks enforce Lagrangian or Hamiltonian structure to improve prediction and generalization. However, when coordinates are embedded in high-dimensional data such as images, these approaches either lose interpretability or can only be applied to one particular example. We introduce a new unsupervised neural network model that learns Lagrangian dynamics from images, with interpretability that benefits prediction and control. The model infers Lagrangian dynamics on generalized coordinates that are simultaneously learned with a coordinate-aware variational autoencoder (VAE). The VAE is designed to account for the geometry of physical systems composed of multiple rigid bodies in the plane. By inferring interpretable Lagrangian dynamics, the model learns physical system properties, such as kinetic and potential energy, which enables long-term prediction of dynamics in the image space and synthesis of energy-based controllers.

[High-Dimensional Sparse Linear Bandits](#)

- Botao Hao, Tor Lattimore, Mengdi Wang
- abstract@[open-review](#): Stochastic linear bandits with high-dimensional sparse features are a practical model for a variety of domains, such as personalized medicine and online advertising. We derive a novel $O(n^{2/3})$ dimension-free minimax regret lower bound for sparse linear bandits in the data-poor regime where the horizon is larger than the ambient dimension and where the feature vectors admit a well-conditioned exploration distribution. This is complemented by a nearly matching upper bound for an explore-then-commit algorithm showing that that $O(n^{2/3})$ is the optimal rate in the data-poor regime. The results complement existing bounds for the data-rich regime and also provide another example where carefully balancing the trade-off between information and regret is necessary. Finally, we prove a dimension-free $O(\sqrt{n})$ regret upper bound under an additional assumption on the magnitude of the signal for relevant features.

[Non-Stochastic Control with Bandit Feedback](#)

- Paula Gradu, John Hallman, Elad Hazan

- abstract@[open-review](#): We study the problem of controlling a linear dynamical system with adversarial perturbations where the only feedback available to the controller is the scalar loss, and the loss function itself is unknown. For this problem, with either a known or unknown system, we give an efficient sublinear regret algorithm. The main algorithmic difficulty is the dependence of the loss on past controls. To overcome this issue, we propose an efficient algorithm for the general setting of bandit convex optimization for loss functions with memory, which may be of independent interest.

[Generalized Leverage Score Sampling for Neural Networks](#)

- Jason D. Lee, Ruoqi Shen, Zhao Song, Mengdi Wang, zheng Yu
- abstract@[open-review](#): Leverage score sampling is a powerful technique that originates from theoretical computer science, which can be used to speed up a large number of fundamental questions, e.g. linear regression, linear programming, semi-definite programming, cutting plane method, graph sparsification, maximum matching and max-flow. Recently, it has been shown that leverage score sampling helps to accelerate kernel methods [Avron, Kapralov, Musco, Musco, Velingker and Zandieh 17]. In this work, we generalize the results in [Avron, Kapralov, Musco, Musco, Velingker and Zandieh 17] to a broader class of kernels. We further bring the leverage score sampling into the field of deep learning theory.
- We show the connection between the initialization for neural network training and approximating the neural tangent kernel with random features.
- We prove the equivalence between regularized neural network and neural tangent kernel ridge regression under the initialization of both classical random Gaussian and leverage score sampling.

[An Optimal Elimination Algorithm for Learning a Best Arm](#)

- Avinatan Hassidim, Ron Kupfer, Yaron Singer
- abstract@[open-review](#): We consider the classic problem of (ϵ, δ) -PAC learning a best arm where the goal is to identify with confidence $1 - \delta$ an arm whose mean is an ϵ -approximation to that of the highest mean arm in a multi-armed bandit setting. This problem is one of the most fundamental problems in statistics and learning theory, yet somewhat surprisingly its worst case sample complexity is not well understood. In this paper we propose a new approach for (ϵ, δ) -PAC learning a best arm. This approach leads to an algorithm whose sample complexity converges to exactly the optimal sample complexity of (ϵ, δ) -learning the mean of n arms separately and we complement this result with a conditional matching lower bound. More specifically:

```
\begin{itemize}
\item The algorithm's sample complexity converges to exactly  $\frac{n}{2\epsilon^2} \log \frac{1}{\delta}$  as  $n$  grows and  $\delta \geq \frac{1}{n}$ ;
\item We prove that no elimination algorithm obtains sample complexity arbitrarily lower than  $\frac{n}{2\epsilon^2} \log \frac{1}{\delta}$ . Elimination algorithms is a broad class of  $(\epsilon, \delta)$ -PAC best arm learning algorithms that includes many algorithms in the literature.
\end{itemize}
When  $n$  is independent of  $\delta$  our approach yields an algorithm whose sample complexity converges to  $\frac{2n}{\epsilon^2} \log \frac{1}{\delta}$  as  $n$  grows. In comparison with the best known algorithm for this problem our approach improves the sample complexity by a factor of over 1500 and over 6000 when  $\delta \geq \frac{1}{n}$ .
```

[Efficient Projection-free Algorithms for Saddle Point Problems](#)

- Cheng Chen, Luo Luo, Weinan Zhang, Yong Yu
- abstract@[open-review](#): The Frank-Wolfe algorithm is a classic method for constrained optimization problems. It has recently been popular in many machine learning applications because its projection-free property leads to more efficient iterations. In this paper, we study projection-free algorithms for convex-strongly-concave saddle point problems with complicated constraints. Our method combines Conditional Gradient Sliding with Mirror-Prox and show that it only requires $\tilde{O}(1/\sqrt{\epsilon})$ gradient evaluations and $\tilde{O}(1/\epsilon^2)$ linear optimizations in the batch setting. We also extend our method to the stochastic setting and propose first stochastic projection-free algorithms for saddle point problems. Experimental results demonstrate the effectiveness of our algorithms and verify our theoretical guarantees.

[A mathematical model for automatic differentiation in machine learning](#)

- Jérôme Bolte, Edouard Pauwels
- abstract@[open-review](#): Automatic differentiation, as implemented today, does not have a simple mathematical model adapted to the needs of modern machine learning. In this work we articulate the relationships between differentiation of programs as implemented in practice, and differentiation of nonsmooth functions. To this end we provide a simple class of functions, a nonsmooth calculus, and show how they apply to stochastic approximation methods. We also evidence the issue of artificial critical points created by algorithmic differentiation and show how usual methods avoid these points with probability one.

[Unsupervised Text Generation by Learning from Search](#)

- Jingjing Li, Zichao Li, Lili Mou, Xin Jiang, Michael Lyu, Irwin King
- abstract@[open-review](#): In this work, we propose TGLS, a novel framework for unsupervised Text Generation by Learning from Search. We start by applying a strong search algorithm (in particular, simulated annealing) towards a heuristically defined objective that (roughly) estimates the quality of sentences. Then, a conditional generative model learns from the search results, and meanwhile smooth out the noise of search. The alternation between search and learning can be repeated for performance bootstrapping. We demonstrate the effectiveness of TGLS on two real-world natural language generation tasks, unsupervised paraphrasing and text formalization. Our model significantly outperforms unsupervised baseline methods in both tasks. Especially, it achieves comparable performance to strong supervised methods for paraphrase generation.

[Learning Compositional Rules via Neural Program Synthesis](#)

- Maxwell Nye, Armando Solar-Lezama, Josh Tenenbaum, Brenden M. Lake
- abstract@[open-review](#): Many aspects of human reasoning, including language, require learning rules from very little data. Humans can do this, often learning systematic rules from very few examples, and combining these rules to form compositional rule-based systems. Current neural architectures, on the other hand, often fail to generalize in a compositional manner, especially when evaluated in ways that vary systematically from training. In this work, we present a neuro-symbolic model which learns entire rule systems from a small set of examples. Instead of directly predicting outputs from inputs, we train our model to induce the explicit system of rules governing a set of previously seen examples, drawing upon techniques from the neural program synthesis literature. Our rule-synthesis approach outperforms neural meta-learning techniques in three domains: an artificial instruction-learning domain used to evaluate human learning, the SCAN challenge datasets, and learning rule-based translations of number words into integers for a wide range of human languages.

[Incorporating BERT into Parallel Sequence Decoding with Adapters](#)

- Junliang Guo, Zhirui Zhang, Linli Xu, Hao-Ran Wei, Boxing Chen, Enhong Chen
- abstract@[open-review](#): While large scale pre-trained language models such as BERT have achieved great success on various natural language understanding tasks, how to efficiently and effectively incorporate them into sequence-to-sequence models and the corresponding text generation tasks remains a non-trivial problem. In this paper, we propose to address this problem by taking two different BERT models as the encoder and decoder

respectively, and fine-tuning them by introducing simple and lightweight adapter modules, which are inserted between BERT layers and tuned on the task-specific dataset. In this way, we obtain a flexible and efficient model which is able to jointly leverage the information contained in the source-side and target-side BERT models, while bypassing the catastrophic forgetting problem. Each component in the framework can be considered as a plug-in unit, making the framework flexible and task agnostic. Our framework is based on a parallel sequence decoding algorithm named Mask-Predict considering the bi-directional and conditional independent nature of BERT, and can be adapted to traditional autoregressive decoding easily. We conduct extensive experiments on neural machine translation tasks where the proposed method consistently outperforms autoregressive baselines while reducing the inference latency by half, and achieves \$36.49\\$/\\$33.57\\$ BLEU scores on IWSLT14 German-English/WMT14 German-English translation. When adapted to autoregressive decoding, the proposed method achieves \$30.60\\$/\\$43.56\\$ BLEU scores on WMT14 English-German/English-French translation, on par with the state-of-the-art baseline models.

[Estimating Fluctuations in Neural Representations of Uncertain Environments](#)

- Sahand Farhoodi, Mark Plitt, Lisa Giocomo, Uri Eden
- abstract@[open-review](#): Neural Coding analyses often reflect an assumption that neural populations respond uniquely and consistently to particular stimuli. For example, analyses of spatial remapping in hippocampal populations often assume that each environment has one unique representation and that remapping occurs over long time scales as an animal traverses between distinct environments. However, as neuroscience experiments begin to explore more naturalistic tasks and stimuli, and reflect more ambiguity in neural representations, methods for analyzing population neural codes must adapt to reflect these features. In this paper, we develop a new state-space modeling framework to address two important issues related to remapping. First, neurons may exhibit significant trial-to-trial or moment-to-moment variability in the firing patterns used to represent a particular environment or stimulus. Second, in ambiguous environments and tasks that involve cognitive uncertainty, neural populations may rapidly fluctuate between multiple representations. The state-space model addresses these two issues by integrating an observation model, which allows for multiple representations of the same stimulus or environment, with a state model, which characterizes the moment-by-moment probability of a shift in the neural representation. These models allow us to compute instantaneous estimates of the stimulus or environment currently represented by the population. We demonstrate the application of this approach to the analysis of population activity in the CA1 region of hippocampus of a mouse moving through ambiguous virtual environments. Our analyses demonstrate that many hippocampal cells express significant trial-to-trial variability in their representations and that the population representation can fluctuate rapidly between environments within a single trial when spatial cues are most ambiguous.

[Discover, Hallucinate, and Adapt: Open Compound Domain Adaptation for Semantic Segmentation](#)

- KwanYong Park, Sanghyun Woo, Inkyu Shin, In So Kweon
- abstract@[open-review](#): Unsupervised domain adaptation (UDA) for semantic segmentation has been attracting attention recently, as it could be beneficial for various label-scarce real-world scenarios (e.g., robot control, autonomous driving, medical imaging, etc.). Despite the significant progress in this field, current works mainly focus on a single-source single-target setting, which cannot handle more practical settings of multiple targets or even unseen targets. In this paper, we investigate open compound domain adaptation (OCDA), which deals with mixed and novel situations at the same time, for semantic segmentation. We present a novel framework based on three main design principles: discover, hallucinate, and adapt. The scheme first clusters compound target data based on style, discovering multiple latent domains (discover). Then, it hallucinates multiple latent target domains in source by using image-translation (hallucinate). This step ensures the latent domains in the source and the target to be paired. Finally, target-to-source alignment is learned separately between domains (adapt). In high-level, our solution replaces a hard OCDA problem with much easier multiple UDA problems. We evaluate our solution on standard benchmark GTA to C-driving, and achieved new state-of-the-art results.

[SURF: A Simple, Universal, Robust, Fast Distribution Learning Algorithm](#)

- Yi Hao, Ayush Jain, Alon Orlitsky, Vaishakh Ravindrakumar
- abstract@[open-review](#): Sample- and computationally-efficient distribution estimation is a fundamental tenet in statistics and machine learning. We present \$\text{SURF}\$, an algorithm for approximating distributions by piecewise polynomials. \$\text{SURF}\$ is: simple, replacing prior complex optimization techniques by straight-forward empirical probability approximation of each potential polynomial piece through simple empirical-probability interpolation, and using plain divide-and-conquer to merge the pieces; universal, as well-known polynomial-approximation results imply that it accurately approximates a large class of common distributions; robust to distribution mis-specification as for any degree \$d \leq 8\$, it estimates any distribution to an \$\ell_1\$ distance \$< 3\$ times that of the nearest degree-\$d\$ piecewise polynomial, improving known factor upper bounds of 3 for single polynomials and 15 for polynomials with arbitrarily many pieces; fast, using optimal sample complexity, running in near sample-linear time, and if given sorted samples it may be parallelized to run in sub-linear time. In experiments, \$\text{SURF}\$ outperforms state-of-the art algorithms.

[Understanding Approximate Fisher Information for Fast Convergence of Natural Gradient Descent in Wide Neural Networks](#)

- Ryo Karakida, Kazuki Osawa
- abstract@[open-review](#): Natural Gradient Descent (NGD) helps to accelerate the convergence of gradient descent dynamics, but it requires approximations in large-scale deep neural networks because of its high computational cost. Empirical studies have confirmed that some NGD methods with approximate Fisher information converge sufficiently fast in practice. Nevertheless, it remains unclear from the theoretical perspective why and under what conditions such heuristic approximations work well. In this work, we reveal that, under specific conditions, NGD with approximate Fisher information achieves the same fast convergence to global minima as exact NGD. We consider deep neural networks in the infinite-width limit, and analyze the asymptotic training dynamics of NGD in function space via the neural tangent kernel. In the function space, the training dynamics with the approximate Fisher information are identical to those with the exact Fisher information, and they converge quickly. The fast convergence holds in layer-wise approximations; for instance, in block diagonal approximation where each block corresponds to a layer as well as in block tri-diagonal and K-FAC approximations. We also find that a unit-wise approximation achieves the same fast convergence under some assumptions. All of these different approximations have an isotropic gradient in the function space, and this plays a fundamental role in achieving the same convergence properties in training. Thus, the current study gives a novel and unified theoretical foundation with which to understand NGD methods in deep learning.

[General Transportability of Soft Interventions: Completeness Results](#)

- Juan Correa, Elias Bareinboim
- abstract@[open-review](#): The challenge of generalizing causal knowledge across different environments is pervasive in scientific explorations, including in AI, ML, and Data Science. Experiments are usually performed in one environment (e.g., in a lab, on Earth) with the intent, almost invariably, of being used elsewhere (e.g., outside the lab, on Mars), where the conditions are likely to be different. In the causal inference literature, this generalization task has been formalized under the rubric of transportability (Pearl and Bareinboim, 2011), where a number of criteria and algorithms have been developed for various settings. Despite the generality of such results, transportability theory has been confined to atomic, do()-interventions. In practice, many real-world applications require more complex, stochastic interventions; for instance, in reinforcement learning, agents need to continuously adapt to the changing conditions of an uncertain and unknown environment.

In this paper, we extend transportability theory to encompass these more complex types of interventions, which are known as "soft," both relative to the input as well as the target distribution of the analysis. Specifically, we develop a graphical condition that is both necessary and sufficient for deciding soft-transportability. Second, we develop an algorithm to determine whether a non-atomic intervention is computable from a combination of the distributions available across domains. As a corollary, we show that the \$\sigma\$-calculus is complete for the task of soft-transportability.

GAIT-prop: A biologically plausible learning rule derived from backpropagation of error

- Nasir Ahmad, Marcel A. J. van Gerven, Luca Ambrogioni
- abstract@[open-review](#): Traditional backpropagation of error, though a highly successful algorithm for learning in artificial neural network models, includes features which are biologically implausible for learning in real neural circuits. An alternative called target propagation proposes to solve this implausibility by using a top-down model of neural activity to convert an error at the output of a neural network into layer-wise and plausible targets™ for every unit. These targets can then be used to produce weight updates for network training. However, thus far, target propagation has been heuristically proposed without demonstrable equivalence to backpropagation. Here, we derive an exact correspondence between backpropagation and a modified form of target propagation (GAIT-prop) where the target is a small perturbation of the forward pass. Specifically, backpropagation and GAIT-prop give identical updates when synaptic weight matrices are orthogonal. In a series of simple computer vision experiments, we show near-identical performance between backpropagation and GAIT-prop with a soft orthogonality-inducing regularizer.

Lipschitz Bounds and Provably Robust Training by Laplacian Smoothing

- Vishaal Krishnan, Abed AlRahman Al Makdah, Fabio Pasqualetti
- abstract@[open-review](#): In this work we propose a graph-based learning framework to train models with provable robustness to adversarial perturbations. In contrast to regularization-based approaches, we formulate the adversarially robust learning problem as one of loss minimization with a Lipschitz constraint, and show that the saddle point of the associated Lagrangian is characterized by a Poisson equation with weighted Laplace operator. Further, the weighting for the Laplace operator is given by the Lagrange multiplier for the Lipschitz constraint, which modulates the sensitivity of the minimizer to perturbations. We then design a provably robust training scheme using graph-based discretization of the input space and a primal-dual algorithm to converge to the Lagrangian's saddle point. Our analysis establishes a novel connection between elliptic operators with constraint-enforced weighting and adversarial learning. We also study the complementary problem of improving the robustness of minimizers with a margin on their loss, formulated as a loss-constrained minimization problem of the Lipschitz constant. We propose a technique to obtain robustified minimizers, and evaluate fundamental Lipschitz lower bounds by approaching Lipschitz constant minimization via a sequence of gradient \$p\$-norm minimization problems. Ultimately, our results show that, for a desired nominal performance, there exists a fundamental lower bound on the sensitivity to adversarial perturbations that depends only on the loss function and the data distribution, and that improvements in robustness beyond this bound can only be made at the expense of nominal performance. Our training schemes provably achieve these bounds both under constraints on performance and~robustness.

SCOP: Scientific Control for Reliable Neural Network Pruning

- Yehui Tang, Yunhe Wang, Yixing Xu, Dacheng Tao, Chunjing XU, Chao Xu, Chang Xu
- abstract@[open-review](#): This paper proposes a reliable neural network pruning algorithm by setting up a scientific control. Existing pruning methods have developed various hypotheses to approximate the importance of filters to the network and then execute filter pruning accordingly. To increase the reliability of the results, we prefer to have a more rigorous research design by including a scientific control group as an essential part to minimize the effect of all factors except the association between the filter and expected network output. Acting as a control group, knockoff feature is generated to mimic the feature map produced by the network filter, but they are conditionally independent of the example label given the real feature map. We theoretically suggest that the knockoff condition can be approximately preserved given the information propagation of network layers. Besides the real feature map on an intermediate layer, the corresponding knockoff feature is brought in as another auxiliary input signal for the subsequent layers. Redundant filters can be discovered in the adversarial process of different features. Through experiments, we demonstrate the superiority of the proposed algorithm over state-of-the-art methods. For example, our method can reduce 57.8% parameters and 60.2% FLOPs of ResNet-101 with only 0.01% top-1 accuracy loss on ImageNet.

Provably Consistent Partial-Label Learning

- Lei Feng, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, Masashi Sugiyama
- abstract@[open-review](#): Partial-label learning (PLL) is a multi-class classification problem, where each training example is associated with a set of candidate labels. Even though many practical PLL methods have been proposed in the last two decades, there lacks a theoretical understanding of the consistency of those methods - none of the PLL methods hitherto possesses a generation process of candidate label sets, and then it is still unclear why such a method works on a specific dataset and when it may fail given a different dataset. In this paper, we propose the first generation model of candidate label sets, and develop two PLL methods that are guaranteed to be provably consistent, i.e., one is risk-consistent and the other is classifier-consistent. Our methods are advantageous, since they are compatible with any deep network or stochastic optimizer. Furthermore, thanks to the generation model, we would be able to answer the two questions above by testing if the generation model matches given candidate label sets. Experiments on benchmark and real-world datasets validate the effectiveness of the proposed generation model and two PLL methods.

Robust, Accurate Stochastic Optimization for Variational Inference

- Akash Kumar Dhaka, Alejandro Catalina, Michael R. Andersen, Måns Magnusson, Jonathan Huggins, Aki Vehtari
- abstract@[open-review](#): We examine the accuracy of black box variational posterior approximations for parametric models in a probabilistic programming context. The performance of these approximations depends on (1) how well the variational family approximates the true posterior distribution, (2) the choice of divergence, and (3) the optimization of the variational objective. We show that even when the true variational family is used, high-dimensional posteriors can be very poorly approximated using common stochastic gradient descent (SGD) optimizers. Motivated by recent theory, we propose a simple and parallel way to improve SGD estimates for variational inference. The approach is theoretically motivated and comes with a diagnostic for convergence and a novel stopping rule, which is robust to noisy objective functions evaluations. We show empirically, the new workflow works well on a diverse set of models and datasets, or warns if the stochastic optimization fails or if the used variational distribution is not good.

Discovering conflicting groups in signed networks

- Ruo-Chun Tzeng, Bruno Ordozoiti, Aristides Gionis
- abstract@[open-review](#): Signed networks are graphs where edges are annotated with a positive or negative sign, indicating whether an edge interaction is friendly or antagonistic. Signed networks can be used to study a variety of social phenomena, such as mining polarized discussions in social media, or modeling relations of trust and distrust in online review platforms. In this paper we study the problem of detecting \$k\$ conflicting groups in a signed network. Our premise is that each group is positively connected internally and negatively connected with the other \$k-1\$ groups. An important aspect of our formulation is that we are not searching for a complete partition of the signed network, instead, we allow other nodes to be neutral with respect to the conflict structure we are searching. As a result, the problem we tackle differs from previously studied problems, such as correlation clustering and \$k\$-way partitioning. To solve the conflicting-group discovery problem, we derive a novel formulation in which each conflicting group is naturally characterized by the solution to the maximum discrete Rayleigh's quotient (\maxdrq) problem. We present two spectral methods for finding approximate solutions to the \maxdrq problem, which we analyze theoretically. Our experimental evaluation shows that, compared to state-of-the-art baselines, our methods find solutions of higher quality, are faster, and recover ground truth conflicting groups with higher accuracy.

Learning Some Popular Gaussian Graphical Models without Condition Number Bounds

- Jonathan Kelner, Frederic Koehler, Raghu Meka, Ankur Moitra
- abstract@[open-review](#): Gaussian Graphical Models (GGMs) have wide-ranging applications in machine learning and the natural and social sciences. In most of the settings in which they are applied, the number of observed samples is much smaller than the dimension and they are assumed to be sparse. While there are a variety of algorithms (e.g. Graphical Lasso, CLIME) that provably recover the graph structure with a logarithmic number of samples, to do so they require various assumptions on the well-conditioning of the precision matrix that are not information-theoretically necessary. Here we give the first fixed polynomial-time algorithms for learning attractive GGMs and walk-summable GGMs with a logarithmic number of samples without any such assumptions. In particular, our algorithms can tolerate strong dependencies among the variables. Our result for structure recovery in walk-summable GGMs is derived from a more general result for efficient sparse linear regression in walk-summable models without any norm dependencies. We complement our results with experiments showing that many existing algorithms fail even in some simple settings where there are long dependency chains. Our algorithms do not.

[Sense and Sensitivity Analysis: Simple Post-Hoc Analysis of Bias Due to Unobserved Confounding](#)

- Victor Veitch, Anisha Zaveri
- abstract@[open-review](#): It is a truth universally acknowledged that an observed association without known mechanism must be in want of a causal estimate. Causal estimates from observational data will be biased in the presence of unobserved confounding. However, we might hope that the influence of unobserved confounders is weak relative to a large estimated effect. The purpose of this paper is to develop Austen plots, a sensitivity analysis tool to aid such judgments by making it easier to reason about potential bias induced by unobserved confounding. We formalize confounding strength in terms of how strongly the unobserved confounding influences treatment assignment and outcome. For a target level of bias, an Austen plot shows the minimum values of treatment and outcome influence required to induce that level of bias. Austen plots generalize the classic sensitivity analysis approach of Imbens [Imb03]. Critically, Austen plots allow any approach for modeling the observed data. We illustrate the tool by assessing biases for several real causal inference problems, using a variety of machine learning approaches for the initial data analysis. Code, demo data, and a tutorial are available at github.com/anishazaveri/austen_plots.

[Mix and Match: An Optimistic Tree-Search Approach for Learning Models from Mixture Distributions](#)

- Matthew Faw, Rajat Sen, Karthikeyan Shanmugam, Constantine Caramanis, Sanjay Shakkottai
- abstract@[open-review](#): We consider a covariate shift problem where one has access to several different training datasets for the same learning problem and a small validation set which possibly differs from all the individual training distributions. The distribution shift is due, in part, to unobserved features in the datasets. The objective, then, is to find the best mixture distribution over the training datasets (with only observed features) such that training a learning algorithm using this mixture has the best validation performance. Our proposed algorithm, Mix&Match, combines stochastic gradient descent (SGD) with optimistic tree search and model re-use (evolving partially trained models with samples from different mixture distributions) over the space of mixtures, for this task. We prove a novel high probability bound on the final SGD iterate without relying on a global gradient norm bound, and use it to show the advantages of model re-use. Additionally, we provide simple regret guarantees for our algorithm with respect to recovering the optimal mixture, given a total budget of SGD evaluations. Finally, we validate our algorithm on two real-world datasets.

[Understanding Double Descent Requires A Fine-Grained Bias-Variance Decomposition](#)

- Ben Adlam, Jeffrey Pennington
- abstract@[open-review](#): Classical learning theory suggests that the optimal generalization performance of a machine learning model should occur at an intermediate model complexity, with simpler models exhibiting high bias and more complex models exhibiting high variance of the predictive function. However, such a simple trade-off does not adequately describe deep learning models that simultaneously attain low bias and variance in the heavily overparameterized regime. A primary obstacle in explaining this behavior is that deep learning algorithms typically involve multiple sources of randomness whose individual contributions are not visible in the total variance. To enable fine-grained analysis, we describe an interpretable, symmetric decomposition of the variance into terms associated with the randomness from sampling, initialization, and the labels. Moreover, we compute the high-dimensional asymptotic behavior of this decomposition for random feature kernel regression, and analyze the strikingly rich phenomenology that arises. We find that the bias decreases monotonically with the network width, but the variance terms exhibit non-monotonic behavior and can diverge at the interpolation boundary, even in the absence of label noise. The divergence is caused by the interaction between sampling and initialization and can therefore be eliminated by marginalizing over samples (i.e. bagging) or over the initial parameters (i.e. ensemble learning).

[VIME: Extending the Success of Self- and Semi-supervised Learning to Tabular Domain](#)

- Jinsung Yoon, Yao Zhang, James Jordon, Mihaela van der Schaar
- abstract@[open-review](#): Self- and semi-supervised learning frameworks have made significant progress in training machine learning models with limited labeled data in image and language domains. These methods heavily rely on the unique structure in the domain datasets (such as spatial relationships in images or semantic relationships in language). They are not adaptable to general tabular data which does not have the same explicit structure as image and language data. In this paper, we fill this gap by proposing novel self- and semi-supervised learning frameworks for tabular data, which we refer to collectively as VIME (Value Imputation and Mask Estimation). We create a novel pretext task of estimating mask vectors from corrupted tabular data in addition to the reconstruction pretext task for self-supervised learning. We also introduce a novel tabular data augmentation method for self- and semi-supervised learning frameworks. In experiments, we evaluate the proposed framework in multiple tabular datasets from various application domains, such as genomics and clinical data. VIME exceeds state-of-the-art performance in comparison to the existing baseline methods.

[The Smoothed Possibility of Social Choice](#)

- Lirong Xia
- abstract@[open-review](#): We develop a framework that leverages the smoothed complexity analysis by Spielman and Teng to circumvent paradoxes and impossibility theorems in social choice, motivated by modern applications of social choice powered by AI and ML. For Condorcet's paradox, we prove that the smoothed likelihood of the paradox either vanishes at an exponential rate as the number of agents increases, or does not vanish at all. For the ANR impossibility on the non-existence of voting rules that simultaneously satisfy anonymity, neutrality, and resolvability, we characterize the rate for the impossibility to vanish, to be either polynomially fast or exponentially fast. We also propose a novel easy-to-compute tie-breaking mechanism that optimally preserves anonymity and neutrality for even number of alternatives in natural settings. Our results illustrate the smoothed possibility of social choice even though the paradox and the impossibility theorem hold in the worst case, they may not be a big concern in practice.

[A Decentralized Parallel Algorithm for Training Generative Adversarial Nets](#)

- Mingrui Liu, Wei Zhang, Youssef Mroueh, Xiaodong Cui, Jarret Ross, Tianbao Yang, Payel Das
- abstract@[open-review](#): Generative Adversarial Networks (GANs) are a powerful class of generative models in the deep learning community. Current practice on large-scale GAN training utilizes large models and distributed large-batch training strategies, and is implemented on deep learning frameworks (e.g., TensorFlow, PyTorch, etc.) designed in a centralized manner. In the centralized network topology, every worker needs to either directly communicate with the central node or indirectly communicate with all other workers in every iteration. However, when the network bandwidth is low or network latency is high, the performance would be significantly degraded. Despite recent progress on decentralized algorithms for training deep neural networks, it remains unclear whether it is possible to train GANs in a decentralized manner. The main difficulty lies at handling the nonconvex-nonconcave min-max

optimization and the decentralized communication simultaneously. In this paper, we address this difficulty by designing the \textbf{first gradient-based decentralized parallel algorithm} which allows workers to have multiple rounds of communications in one iteration and to update the discriminator and generator simultaneously, and this design makes it amenable for the convergence analysis of the proposed decentralized algorithm. Theoretically, our proposed decentralized algorithm is able to solve a class of non-convex non-concave min-max problems with provable non-asymptotic convergence to first-order stationary point. Experimental results on GANs demonstrate the effectiveness of the proposed algorithm.

[Phase retrieval in high dimensions: Statistical and computational phase transitions](#)

- Antoine Maillard, Bruno Loureiro, Florent Krzakala, Lenka Zdeborová;
- abstract@[open-review](#): We consider the phase retrieval problem of reconstructing a n -dimensional real or complex signal \mathbf{X}^* from m (possibly noisy) observations $\mathbf{Y}_\mu = \sum_{i=1}^n \Phi_i \mu_i \mathbf{X}^* \sqrt{n}$, for a large class of correlated real and complex random sensing matrices Φ , in a high-dimensional setting where $m, n \rightarrow \infty$ while $\alpha = m/n = \Theta(1)$. First, we derive sharp asymptotics for the lowest possible estimation error achievable statistically and we unveil the existence of sharp phase transitions for the weak- and full-recovery thresholds as a function of the singular values of the matrix Φ . This is achieved by providing a rigorous proof of a result first obtained by the replica method from statistical mechanics. In particular, the information-theoretic transition to perfect recovery for full-rank matrices appears at $\alpha=1$ (real case) and $\alpha=2$ (complex case). Secondly, we analyze the performance of the best-known polynomial time algorithm for this problem --- approximate message-passing--- establishing the existence of statistical-to-algorithmic gap depending, again, on the spectral properties of Φ . Our work provides an extensive classification of the statistical and algorithmic thresholds in high-dimensional phase retrieval for a broad class of random matrices.

[Fair Performance Metric Elicitation](#)

- Gaurush Hiranandani, Harikrishna Narasimhan, Sanmi Koyejo
- abstract@[open-review](#): What is a fair performance metric? We consider the choice of fairness metrics through the lens of metric elicitation -- a principled framework for selecting performance metrics that best reflect implicit preferences. The use of metric elicitation enables a practitioner to tune the performance and fairness metrics to the task, context, and population at hand. Specifically, we propose a novel strategy to elicit group-fair performance metrics for multiclass classification problems with multiple sensitive groups that also includes selecting the trade-off between predictive performance and fairness violation. The proposed elicitation strategy requires only relative preference feedback and is robust to both finite sample and feedback noise.

[Hybrid Variance-Reduced SGD Algorithms For Minimax Problems with Nonconvex-Linear Function](#)

- Quoc Tran Dinh, Deyi Liu, Lam Nguyen
- abstract@[open-review](#): We develop a novel and single-loop variance-reduced algorithm to solve a class of stochastic nonconvex-convex minimax problems involving a nonconvex-linear objective function, which has various applications in different fields such as machine learning and robust optimization. This problem class has several computational challenges due to its nonsmoothness, nonconvexity, nonlinearity, and non-separability of the objective functions. Our approach relies on a new combination of recent ideas, including smoothing and hybrid biased variance-reduced techniques. Our algorithm and its variants can achieve $O(T^{-2/3})$ -convergence rate and the best-known oracle complexity under standard assumptions, where T is the iteration counter. They have several computational advantages compared to existing methods such as simple to implement and less parameter tuning requirements. They can also work with both single sample or mini-batch on derivative estimators, and with constant or diminishing step-sizes. We demonstrate the benefits of our algorithms over existing methods through two numerical examples, including a nonsmooth and nonconvex-non-strongly concave minimax model.

[Belief-Dependent Macro-Action Discovery in POMDPs using the Value of Information](#)

- Genevieve Flaspohler, Nicholas A. Roy, John W. Fisher III
- abstract@[open-review](#): This work introduces macro-action discovery using value-of-information (VoI) for robust and efficient planning in partially observable Markov decision processes (POMDPs). POMDPs are a powerful framework for planning under uncertainty. Previous approaches have used high-level macro-actions within POMDP policies to reduce planning complexity. However, macro-action design is often heuristic and rarely comes with performance guarantees. Here, we present a method for extracting belief-dependent, variable-length macro-actions directly from a low-level POMDP model. We construct macro-actions by chaining sequences of open-loop actions together when the task-specific value of information (VoI) --- the change in expected task performance caused by observations in the current planning iteration --- is low. Importantly, we provide performance guarantees on the resulting VoI macro-action policies in the form of bounded regret relative to the optimal policy. In simulated tracking experiments, we achieve higher reward than both closed-loop and hand-coded macro-action baselines, selectively using VoI macro-actions to reduce planning complexity while maintaining near-optimal task performance.

[Soft Contrastive Learning for Visual Localization](#)

- Janine Thoma, Danda Pani Paudel, Luc V. Gool
- abstract@[open-review](#): Localization by image retrieval is inexpensive and scalable due to simple mapping and matching techniques. Such localization, however, depends upon the quality of image features often obtained using Contrastive learning frameworks. Most contrastive learning strategies opt for features to distinguish different classes. In the context of localization, however, there is no natural definition of classes. Therefore, images are usually artificially separated into positive and negative classes, with respect to the chosen anchor images, based on some geometric proximity measure. In this paper, we show why such divisions are problematic for learning localization features. We argue that any artificial division based on some proximity measure is undesirable, due to the inherently ambiguous supervision for images near proximity threshold. To this end, we propose a novel technique that uses soft positive/negative assignments of images for contrastive learning, avoiding the aforementioned problem. Our soft assignment makes a gradual distinction between close and far images in both geometric and feature spaces. Experiments on four large-scale benchmark datasets demonstrate the superiority of the proposed soft contrastive learning over the state-of-the-art method for retrieval-based visual localization.

[Fine-Grained Dynamic Head for Object Detection](#)

- Lin Song, Yanwei Li, Zhengkai Jiang, Zeming Li, Hongbin Sun, Jian Sun, Nanning Zheng
- abstract@[open-review](#): The Feature Pyramid Network (FPN) presents a remarkable approach to alleviate the scale variance in object representation by performing instance-level assignments. Nevertheless, this strategy ignores the distinct characteristics of different sub-regions in an instance. To this end, we propose a fine-grained dynamic head to conditionally select a pixel-level combination of FPN features from different scales for each instance, which further releases the ability of multi-scale feature representation. Moreover, we design a spatial gate with the new activation function to reduce computational complexity dramatically through spatially sparse convolutions. Extensive experiments demonstrate the effectiveness and efficiency of the proposed method on several state-of-the-art detection benchmarks. Code is available at <https://github.com/StevenGrove/DynamicHead>.

[LoCo: Local Contrastive Representation Learning](#)

- Yuwen Xiong, Mengye Ren, Raquel Urtasun

- abstract@[open-review](#): Deep neural nets typically perform end-to-end backpropagation to learn the weights, a procedure that creates synchronization constraints in the weight update step across layers and is not biologically plausible. Recent advances in unsupervised contrastive representation learning invite the question of whether a learning algorithm can also be made local, that is, the updates of lower layers do not directly depend on the computation of upper layers. While Greedy InfoMax separately learns each block with a local objective, we found that it consistently hurts readout accuracy in state-of-the-art unsupervised contrastive learning algorithms, possibly due to the greedy objective as well as gradient isolation. In this work, we discover that by overlapping local blocks stacking on top of each other, we effectively increase the decoder depth and allow upper blocks to implicitly send feedbacks to lower blocks. This simple design closes the performance gap between local learning and end-to-end contrastive learning algorithms for the first time. Aside from standard ImageNet experiments, we also show results on complex downstream tasks such as object detection and instance segmentation directly using readout features.

[Modeling and Optimization Trade-off in Meta-learning](#)

- Katelyn Gao, Ozan Sener
- abstract@[open-review](#): By searching for shared inductive biases across tasks, meta-learning promises to accelerate learning on novel tasks, but with the cost of solving a complex bilevel optimization problem. We introduce and rigorously define the trade-off between accurate modeling and optimization ease in meta-learning. At one end, classic meta-learning algorithms account for the structure of meta-learning but solve a complex optimization problem, while at the other end domain randomized search (otherwise known as joint training) ignores the structure of meta-learning and solves a single level optimization problem. Taking MAML as the representative meta-learning algorithm, we theoretically characterize the trade-off for general non-convex risk functions as well as linear regression, for which we are able to provide explicit bounds on the errors associated with modeling and optimization. We also empirically study this trade-off for meta-reinforcement learning benchmarks.

[SnapBoost: A Heterogeneous Boosting Machine](#)

- Thomas Parnell, Andreea Anghel, MaÅ,gorzata Åazuka, Nikolas Ioannou, Sebastian Kurella, Peshal Agarwal, Nikolaos Papandreou, Haralampos Pozidis
- abstract@[open-review](#): Modern gradient boosting software frameworks, such as XGBoost and LightGBM, implement Newton descent in a functional space. At each boosting iteration, their goal is to find the base hypothesis, selected from some base hypothesis class, that is closest to the Newton descent direction in a Euclidean sense. Typically, the base hypothesis class is fixed to be all binary decision trees up to a given depth. In this work, we study a Heterogeneous Newton Boosting Machine (HNBM) in which the base hypothesis class may vary across boosting iterations. Specifically, at each boosting iteration, the base hypothesis class is chosen, from a fixed set of subclasses, by sampling from a probability distribution. We derive a global linear convergence rate for the HNBM under certain assumptions, and show that it agrees with existing rates for Newton's method when the Newton direction can be perfectly fitted by the base hypothesis at each boosting iteration. We then describe a particular realization of a HNBM, SnapBoost, that, at each boosting iteration, randomly selects between either a decision tree of variable depth or a linear regressor with random Fourier features. We describe how SnapBoost is implemented, with a focus on the training complexity. Finally, we present experimental results, using OpenML and Kaggle datasets, that show that SnapBoost is able to achieve better generalization loss than competing boosting frameworks, without taking significantly longer to tune.

[On Adaptive Distance Estimation](#)

- Yeshwanth Cherapanamjeri, Jelani Nelson
- abstract@[open-review](#): We provide a static data structure for distance estimation which supports {\it adaptive} queries. Concretely, given a dataset $X = \{x_i\}_{i=1}^n$ of n points in \mathbb{R}^d and $0 < p \leq 2$, we construct a randomized data structure with low memory consumption and query time which, when later given any query point $q \in \mathbb{R}^d$, outputs a $(1+\epsilon)$ -approximation of $\|q - x_i\|_p$ with high probability for all $i \in [n]$. The main novelty is our data structure's correctness guarantee holds even when the sequence of queries can be chosen adaptively: an adversary is allowed to choose the j th query point q_j in a way that depends on the answers reported by the data structure for q_1, \dots, q_{j-1} . Previous randomized Monte Carlo methods do not provide error guarantees in the setting of adaptively chosen queries. Our memory consumption is $\tilde{O}(nd/\epsilon^2)$, slightly more than the $O(nd)$ required to store X in memory explicitly, but with the benefit that our time to answer queries is only $\tilde{O}(\epsilon^{-2}(n+d))$, much faster than the naive $\Theta(nd)$ time obtained from a linear scan in the case of n and d very large. Here \tilde{O} hides $\log(nd/\epsilon)$ factors. We discuss applications to nearest neighbor search and nonparametric estimation.

Our method is simple and likely to applicable to other domains: we describe a generic approach for transforming randomized Monte Carlo data structures which do not support adaptive queries to ones that do, and show that for the problem at hand it can be applied to standard nonadaptive solutions to ℓ_p norm estimation with negligible overhead in query time and a factor d overhead in memory.

[Stage-wise Conservative Linear Bandits](#)

- Ahmadreza Moradipari, Christos Thrampoulidis, Mahnoosh Alizadeh
- abstract@[open-review](#): We study stage-wise conservative linear stochastic bandits: an instance of bandit optimization, which accounts for (unknown) safety constraints that appear in applications such as online advertising and medical trials. At each stage, the learner must choose actions that not only maximize cumulative reward across the entire time horizon, but further satisfy a linear baseline constraint that takes the form of a lower bound on the instantaneous reward. For this problem, we present two novel algorithms, stage-wise conservative linear Thompson Sampling (SCLTS) and stage-wise conservative linear UCB (SCLUCB), that respect the baseline constraints and enjoy probabilistic regret bounds of order $\mathcal{O}(\sqrt{T})\log^{3/2}T$ and $\mathcal{O}(\sqrt{T}\log T)$, respectively. Notably, the proposed algorithms can be adjusted with only minor modifications to tackle different problem variations, such as, constraints with bandit-feedback, or an unknown sequence of baseline rewards. We discuss these and other improvements over the state-of-the art. For instance, compared to existing solutions, we show that SCLTS plays the (non-optimal) baseline action at most $\mathcal{O}(\log T)$ times (compared to $\mathcal{O}(\sqrt{T})$). Finally, we make connections to another studied form of safety-constraints that takes the form of an upper bound on the instantaneous reward. While this incurs additional complexity to the learning process as the optimal action is not guaranteed to belong to the safe-set at each round, we show that SCLUCB can properly adjust in this setting via a simple modification.

[RELATE: Physically Plausible Multi-Object Scene Synthesis Using Structured Latent Spaces](#)

- Sebastien Ehrhardt, Oliver Groth, Aron Monszpart, Martin Engelcke, Ingmar Posner, Niloy Mitra, Andrea Vedaldi
- abstract@[open-review](#): We present RELATE, a model that learns to generate physically plausible scenes and videos of multiple interacting objects. Similar to other generative approaches, RELATE is trained end-to-end on raw, unlabeled data. RELATE combines an object-centric GAN formulation with a model that explicitly accounts for correlations between individual objects. This allows the model to generate realistic scenes and videos from a physically-interpretable parameterization. Furthermore, we show that modeling the object correlation is necessary to learn to disentangle object positions and identity. We find that RELATE is also amenable to physically realistic scene editing and that it significantly outperforms prior art in object-centric scene generation in both synthetic (CLEVR, ShapeStacks) and real-world data (cars). In addition, in contrast to state-of-the-art methods in object-centric generative modeling, RELATE also extends naturally to dynamic scenes and generates videos of high visual fidelity. Source code, datasets and more results are available at <http://geometry.cs.ucl.ac.uk/projects/2020/relate/>.

[Metric-Free Individual Fairness in Online Learning](#)

- Yahav Bechavod, Christopher Jung, Steven Z. Wu
- abstract@[open-review](#): We study an online learning problem subject to the constraint of individual fairness, which requires that similar individuals are treated similarly. Unlike prior work on individual fairness, we do not assume the similarity measure among individuals is known, nor do we assume that such measure takes a certain parametric form. Instead, we leverage the existence of an auditor who detects fairness violations without enunciating the quantitative measure. In each round, the auditor examines the learner's decisions and attempts to identify a pair of individuals that are treated unfairly by the learner. We provide a general reduction framework that reduces online classification in our model to standard online classification, which allows us to leverage existing online learning algorithms to achieve sub-linear regret and number of fairness violations. Surprisingly, in the stochastic setting where the data are drawn independently from a distribution, we are also able to establish PAC-style fairness and accuracy generalization guarantees (Rothblum and Yona (2018)), despite only having access to a very restricted form of fairness feedback. Our fairness generalization bound qualitatively matches the uniform convergence bound of Rothblum and Yona (2018), while also providing a meaningful accuracy generalization guarantee. Our results resolve an open question by Gillen et al. (2018) by showing that online learning under an unknown individual fairness constraint is possible even without assuming a strong parametric form of the underlying similarity measure.

[GreedyFool: Distortion-Aware Sparse Adversarial Attack](#)

- Xiaoyi Dong, Dongdong Chen, Jianmin Bao, Chuan Qin, Lu Yuan, Weiming Zhang, Nenghai Yu, Dong Chen
- abstract@[open-review](#): Modern deep neural networks(DNNs) are vulnerable to adversarial samples. Sparse adversarial samples are a special branch of adversarial samples that can fool the target model by only perturbing a few pixels. The existence of the sparse adversarial attack points out that DNNs are much more vulnerable than people believed, which is also a new aspect for analyzing DNNs. However, current sparse adversarial attack methods still have some shortcomings on both sparsity and invisibility. In this paper, we propose a novel two-stage distortion-aware greedy-based method dubbed as "GreedyFool". Specifically, it first selects the most effective candidate positions to modify by considering both the gradient(for adversary) and the distortion map(for invisibility), then drops some less important points in the reduce stage. Experiments demonstrate that compared with the start-of-the-art method, we only need to modify 3 times fewer pixels under the same sparse perturbation setting. For target attack, the success rate of our method is 9.96% higher than the start-of-the-art method under the same pixel budget.

[VAEM: a Deep Generative Model for Heterogeneous Mixed Type Data](#)

- Chao Ma, Sebastian Tschiatschek, Richard Turner, JosÃ© Miguel HernÃ¡ndez-Lobato, Cheng Zhang
- abstract@[open-review](#): Deep generative models often perform poorly in real-world applications due to the heterogeneity of natural data sets. Heterogeneity arises from data containing different types of features (categorical, ordinal, continuous, etc.) and features of the same type having different marginal distributions. We propose an extension of variational autoencoders (VAEs) called VAEM to handle such heterogeneous data. VAEM is a deep generative model that is trained in a two stage manner, such that the first stage provides a more uniform representation of the data to the second stage, thereby sidestepping the problems caused by heterogeneous data. We provide extensions of VAEM to handle partially observed data, and demonstrate its performance in data generation, missing data prediction and sequential feature selection tasks. Our results show that VAEM broadens the range of real-world applications where deep generative models can be successfully deployed.

[RetroXpert: Decompose Retrosynthesis Prediction Like A Chemist](#)

- Chaochao Yan, Qianggang Ding, Peilin Zhao, Shuangjia Zheng, JINYU YANG, Yang Yu, Junzhou Huang
- abstract@[open-review](#): Retrosynthesis is the process of recursively decomposing target molecules into available building blocks. It plays an important role in solving problems in organic synthesis planning. To automate or assist in the retrosynthesis analysis, various retrosynthesis prediction algorithms have been proposed. However, most of them are cumbersome and lack interpretability about their predictions. In this paper, we devise a novel template-free algorithm for automatic retrosynthetic expansion inspired by how chemists approach retrosynthesis prediction. Our method disassembles retrosynthesis into two steps: i) identify the potential reaction center of the target molecule through a novel graph neural network and generate intermediate synthons, and ii) generate the reactants associated with synthons via a robust reactant generation model. While outperforming the state-of-the-art baselines by a significant margin, our model also provides chemically reasonable interpretation.

[Sample-Efficient Optimization in the Latent Space of Deep Generative Models via Weighted Retraining](#)

- Austin Tripp, Erik Daxberger, JosÃ© Miguel HernÃ¡ndez-Lobato
- abstract@[open-review](#): Many important problems in science and engineering, such as drug design, involve optimizing an expensive black-box objective function over a complex, high-dimensional, and structured input space. Although machine learning techniques have shown promise in solving such problems, existing approaches substantially lack sample efficiency. We introduce an improved method for efficient black-box optimization, which performs the optimization in the low-dimensional, continuous latent manifold learned by a deep generative model. In contrast to previous approaches, we actively steer the generative model to maintain a latent manifold that is highly useful for efficiently optimizing the objective. We achieve this by periodically retraining the generative model on the data points queried along the optimization trajectory, as well as weighting those data points according to their objective function value. This weighted retraining can be easily implemented on top of existing methods, and is empirically shown to significantly improve their efficiency and performance on synthetic and real-world optimization problems.

[Improved Sample Complexity for Incremental Autonomous Exploration in MDPs](#)

- Jean Tarbouriech, Matteo Pirotta, Michal Valko, Alessandro Lazaric
- abstract@[open-review](#): We study the problem of exploring an unknown environment when no reward function is provided to the agent. Building on the incremental exploration setting introduced by Lim and Auer (2012), we define the objective of learning the set of $\$epsilon$ -optimal goal-conditioned policies attaining all states that are incrementally reachable within $\$L$ steps (in expectation) from a reference state $\$s_0$. In this paper, we introduce a novel model-based approach that interleaves discovering new states from $\$s_0$ and improving the accuracy of a model estimate that is used to compute goal-conditioned policies. The resulting algorithm, DisCo, achieves a sample complexity scaling as $\widetilde{O}(\epsilon^2 L^5 S(L+\epsilon)^2 \Gamma^{L+\epsilon})$, where A is the number of actions, S is the number of states that are incrementally reachable from $\$s_0$ in $L+\epsilon$ steps, and Γ is the branching factor of the dynamics over such states. This improves over the algorithm proposed in (Lim and Auer, 2012) in both ϵ and L at the cost of an extra $\Gamma^{L+\epsilon}$ factor, which is small in most environments of interest. Furthermore, DisCo is the first algorithm that can return an ϵ/c_{\min} -optimal policy for any cost-sensitive shortest-path problem defined on the L -reachable states with minimum cost c_{\min} . Finally, we report preliminary empirical results confirming our theoretical findings.

[TinyTL: Reduce Memory, Not Parameters for Efficient On-Device Learning](#)

- Han Cai, Chuang Gan, Ligeng Zhu, Song Han
- abstract@[open-review](#): Efficient on-device learning requires a small memory footprint at training time to fit the tight memory constraint. Existing work solves this problem by reducing the number of trainable parameters. However, this doesn't directly translate to memory saving since the major bottleneck is the activations, not parameters. In this work, we present Tiny-Transfer-Learning (TinyTL) for memory-efficient on-device learning. TinyTL freezes the weights while only learns the memory-efficient bias modules, thus no need to store the intermediate activations. To maintain the adaptation capacity, we introduce a new memory-efficient bias module, the lite residual module, to refine the feature extractor by learning small residual feature maps adding only

3.8% memory overhead. Extensive experiments show that TinyTL significantly saves the memory (up to 6.5x) with little accuracy loss compared to fine-tuning the full network. Compared to fine-tuning the last layer, TinyTL provides significant accuracy improvements (up to 33.8%) with little memory overhead. Furthermore, combined with feature extractor adaptation, TinyTL provides 7.5-12.9x memory saving without sacrificing accuracy compared to fine-tuning the full Inception-V3. Code is released at <https://github.com/mit-han-lab/tinyML/tree/master/tinyTL>.

[RD\\$^2\\$: Reward Decomposition with Representation Decomposition](#)

- Zichuan Lin, Derek Yang, Li Zhao, Tao Qin, Guangwen Yang, Tie-Yan Liu
- abstract@[open-review](#): Reward decomposition, which aims to decompose the full reward into multiple sub-rewards, has been proven beneficial for improving sample efficiency in reinforcement learning. Existing works on discovering reward decomposition are mostly policy dependent, which constrains diverse or disentangled behavior between different policies induced by different sub-rewards. In this work, we propose a set of novel reward decomposition principles by constraining uniqueness and compactness of different state features/representations relevant to different sub-rewards. Our principles encourage sub-rewards with minimal relevant features, while maintaining the uniqueness of each sub-reward. We derive a deep learning algorithm based on our principle, and term our method as RD\$^2\$, since we learn reward decomposition and representation decomposition jointly. RD\$^2\$ is evaluated on a toy case, where we have the true reward structure, and some Atari environments where reward structure exists but is unknown to the agent to demonstrate the effectiveness of RD\$^2\$ against existing reward decomposition methods.

[Self-paced Contrastive Learning with Hybrid Memory for Domain Adaptive Object Re-ID](#)

- Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, hongsheng Li
- abstract@[open-review](#): Domain adaptive object re-ID aims to transfer the learned knowledge from the labeled source domain to the unlabeled target domain to tackle the open-class re-identification problems. Although state-of-the-art pseudo-label-based methods have achieved great success, they did not make full use of all valuable information because of the domain gap and unsatisfying clustering performance. To solve these problems, we propose a novel self-paced contrastive learning framework with hybrid memory. The hybrid memory dynamically generates source-domain class-level, target-domain cluster-level and un-clustered instance-level supervisory signals for learning feature representations. Different from the conventional contrastive learning strategy, the proposed framework jointly distinguishes source-domain classes, and target-domain clusters and un-clustered instances. Most importantly, the proposed self-paced method gradually creates more reliable clusters to refine the hybrid memory and learning targets, and is shown to be the key to our outstanding performance. Our method outperforms state-of-the-arts on multiple domain adaptation tasks of object re-ID and even boosts the performance on the source domain without any extra annotations. Our generalized version on unsupervised object re-ID surpasses state-of-the-art algorithms by considerable 16.7% and 7.9% on Market-1501 and MSMT17 benchmarks.

[Fairness constraints can help exact inference in structured prediction](#)

- Kevin Bello, Jean Honorio
- abstract@[open-review](#): Many inference problems in structured prediction can be modeled as maximizing a score function on a space of labels, where graphs are a natural representation to decompose the total score into a sum of unary (nodes) and pairwise (edges) scores. Given a generative model with an undirected connected graph G and true vector of binary labels \bar{y} , it has been previously shown that when G has good expansion properties, such as complete graphs or d -regular expanders, one can exactly recover \bar{y} (with high probability and in polynomial time) from a single noisy observation of each edge and node. We analyze the previously studied generative model by Globerson et al. (2015) under a notion of statistical parity. That is, given a fair binary node labeling, we ask the question whether it is possible to recover the fair assignment, with high probability and in polynomial time, from single edge and node observations. We find that, in contrast to the known trade-offs between fairness and model performance, the addition of the fairness constraint improves the probability of exact recovery. We effectively explain this phenomenon and empirically show how graphs with poor expansion properties, such as grids, are now capable of achieving exact recovery. Finally, as a byproduct of our analysis, we provide a tighter minimum-eigenvalue bound than that which can be derived from Weyl's inequality.

[Instance-based Generalization in Reinforcement Learning](#)

- Martin Bertran, Natalia Martinez, Mariano Philipp, Guillermo Sapiro
- abstract@[open-review](#): Agents trained via deep reinforcement learning (RL) routinely fail to generalize to unseen environments, even when these share the same underlying dynamics as the training levels. Understanding the generalization properties of RL is one of the challenges of modern machine learning. Towards this goal, we analyze policy learning in the context of Partially Observable Markov Decision Processes (POMDPs) and formalize the dynamics of training levels as instances. We prove that, independently of the exploration strategy, reusing instances introduces significant changes on the effective Markov dynamics the agent observes during training. Maximizing expected rewards impacts the learned belief state of the agent by inducing undesired instance-specific speed-running policies instead of generalizable ones, which are sub-optimal on the training set. We provide generalization bounds to the value gap in train and test environments based on the number of training instances, and use insights based on these to improve performance on unseen levels. We propose training a shared belief representation over an ensemble of specialized policies, from which we compute a consensus policy that is used for data collection, disallowing instance-specific exploitation. We experimentally validate our theory, observations, and the proposed computational solution over the CoinRun benchmark.

[Smooth And Consistent Probabilistic Regression Trees](#)

- Sami Alkhouri, Emilie Devijver, Marianne Clausel, Myriam Tami, Eric Gaussier, georges Oppenheim
- abstract@[open-review](#): We propose here a generalization of regression trees, referred to as Probabilistic Regression (PR) trees, that adapt to the smoothness of the prediction function relating input and output variables while preserving the interpretability of the prediction and being robust to noise. In PR trees, an observation is associated to all regions of a tree through a probability distribution that reflects how far the observation is to a region. We show that such trees are consistent, meaning that their error tends to 0 when the sample size tends to infinity, a property that has not been established for similar, previous proposals as Soft trees and Smooth Transition Regression trees. We further explain how PR trees can be used in different ensemble methods, namely Random Forests and Gradient Boosted Trees. Lastly, we assess their performance through extensive experiments that illustrate their benefits in terms of performance, interpretability and robustness to noise.

[Computing Valid p-value for Optimal Changepoint by Selective Inference using Dynamic Programming](#)

- Vo Nguyen Le Duy, Hiroki Toda, Ryota Sugiyama, Ichiro Takeuchi
- abstract@[open-review](#): Although there is a vast body of literature related to methods for detecting change-points (CPs), less attention has been paid to assessing the statistical reliability of the detected CPs. In this paper, we introduce a novel method to perform statistical inference on the significance of the CPs, estimated by a Dynamic Programming (DP)-based optimal CP detection algorithm. Our main idea is to employ a Selective Inference (SI) approach---a new statistical inference framework that has recently received a lot of attention---to compute exact (non-asymptotic) valid p-values for the detected optimal CPs. Although it is well-known that SI has low statistical power because of over-conditioning, we address this drawback by introducing a novel method called parametric DP, which enables SI to be conducted with the minimum amount of conditioning, leading to high statistical power. We conduct experiments on both synthetic and real-world datasets, through which we offer evidence that our proposed method is more powerful than existing methods, has decent performance in terms of computational efficiency, and provides good results in many practical applications.

[Factorized Neural Processes for Neural Processes: K-Shot Prediction of Neural Responses](#)

- Ronald (James) Cotton, Fabian Sinz, Andreas Tolias
- abstract@[open-review](#): In recent years, artificial neural networks have achieved state-of-the-art performance for predicting the responses of neurons in the visual cortex to natural stimuli. However, they require a time consuming parameter optimization process for accurately modeling the tuning function of newly observed neurons, which prohibits many applications including real-time, closed-loop experiments. We overcome this limitation by formulating the problem as $K\$$ -shot prediction to directly infer a neuron's tuning function from a small set of stimulus-response pairs using a Neural Process. This required us to developed a Factorized Neural Process, which embeds the observed set into a latent space partitioned into the receptive field location and the tuning function properties. We show on simulated responses that the predictions and reconstructed receptive fields from the Factorized Neural Process approach ground truth with increasing number of trials. Critically, the latent representation that summarizes the tuning function of a neuron is inferred in a quick, single forward pass through the network. Finally, we validate this approach on real neural data from visual cortex and find that the predictive accuracy is comparable to --- and for small $K\$$ even greater than --- optimization based approaches, while being substantially faster. We believe this novel deep learning systems identification framework will facilitate better real-time integration of artificial neural network modeling into neuroscience experiments.

[Winning the Lottery with Continuous Sparsification](#)

- Pedro Savarese, Hugo Silva, Michael Maire
- abstract@[open-review](#): The search for efficient, sparse deep neural network models is most prominently performed by pruning: training a dense, overparameterized network and removing parameters, usually via following a manually-crafted heuristic. Additionally, the recent Lottery Ticket Hypothesis conjectures that, for a typically-sized neural network, it is possible to find small sub-networks which, when trained from scratch on a comparable budget, match the performance of the original dense counterpart. We revisit fundamental aspects of pruning algorithms, pointing out missing ingredients in previous approaches, and develop a method, Continuous Sparsification, which searches for sparse networks based on a novel approximation of an intractable ℓ_0 regularization. We compare against dominant heuristic-based methods on pruning as well as ticket search -- finding sparse subnetworks that can be successfully re-trained from an early iterate. Empirical results show that we surpass the state-of-the-art for both objectives, across models and datasets, including VGG trained on CIFAR-10 and ResNet-50 trained on ImageNet. In addition to setting a new standard for pruning, Continuous Sparsification also offers fast parallel ticket search, opening doors to new applications of the Lottery Ticket Hypothesis.

[Adversarial robustness via robust low rank representations](#)

- Pranjal Awasthi, Himanshu Jain, Ankit Singh Rawat, Aravindan Vijayaraghavan
- abstract@[open-review](#): Adversarial robustness measures the susceptibility of a classifier to imperceptible perturbations made to the inputs at test time. In this work we highlight the benefits of natural low rank representations that often exist for real data such as images, for training neural networks with certified robustness guarantees. Our first contribution is for certified robustness to perturbations measured in L_2 norm. We exploit low rank data representations to provide improved guarantees over state-of-the-art randomized smoothing-based approaches on standard benchmark datasets such as CIFAR-10 and CIFAR-100. Our second contribution is for the more challenging setting of certified robustness to perturbations measured in L_∞ norm. We demonstrate empirically that natural low rank representations have inherent robustness properties that can be leveraged to provide significantly better guarantees for certified robustness to L_∞ perturbations in those representations. Our certificate of L_∞ robustness relies on a natural quantity involving the $\infty \rightarrow 2$ matrix operator norm associated with the representation, to translate robustness guarantees from L_2 to L_∞ perturbations. A key technical ingredient for our certification guarantees is a fast algorithm with provable guarantees based on the multiplicative weights update method to provide upper bounds on the above matrix norm. Our algorithmic guarantees improve upon the state of the art for this problem, and may be of independent interest.

[Joints in Random Forests](#)

- Alvaro Correia, Robert Peharz, Cassio P. de Campos
- abstract@[open-review](#): Decision Trees (DTs) and Random Forests (RFs) are powerful discriminative learners and tools of central importance to the everyday machine learning practitioner and data scientist. Due to their discriminative nature, however, they lack principled methods to process inputs with missing features or to detect outliers, which requires pairing them with imputation techniques or a separate generative model. In this paper, we demonstrate that DTs and RFs can naturally be interpreted as generative models, by drawing a connection to Probabilistic Circuits, a prominent class of tractable probabilistic models. This reinterpretation equips them with a full joint distribution over the feature space and leads to Generative Decision Trees (GeDTs) and Generative Forests (GeFs), a family of novel hybrid generative-discriminative models. This family of models retains the overall characteristics of DTs and RFs while additionally being able to handle missing features by means of marginalisation. Under certain assumptions, frequently made for Bayes consistency results, we show that consistency in GeDTs and GeFs extend to any pattern of missing input features, if missing at random. Empirically, we show that our models often outperform common routines to treat missing data, such as K-nearest neighbour imputation, and moreover, that our models can naturally detect outliers by monitoring the marginal probability of input features.

[Compositional Generalization by Learning Analytical Expressions](#)

- Qian Liu, Shengnan An, Jian-Guang Lou, Bei Chen, Zeqi Lin, Yan Gao, Bin Zhou, Nanning Zheng, Dongmei Zhang
- abstract@[open-review](#): Compositional generalization is a basic and essential intellective capability of human beings, which allows us to recombine known parts readily. However, existing neural network based models have been proven to be extremely deficient in such a capability. Inspired by work in cognition which argues compositionality can be captured by variable slots with symbolic functions, we present a refreshing view that connects a memory-augmented neural model with analytical expressions, to achieve compositional generalization. Our model consists of two cooperative neural modules, Composer and Solver, fitting well with the cognitive argument while being able to be trained in an end-to-end manner via a hierarchical reinforcement learning algorithm. Experiments on the well-known benchmark SCAN demonstrate that our model seizes a great ability of compositional generalization, solving all challenges addressed by previous works with 100% accuracies.

[JAX MD: A Framework for Differentiable Physics](#)

- Samuel Schoenholz, Ekin Dogus Cubuk
- abstract@[open-review](#): We introduce JAX MD, a software package for performing differentiable physics simulations with a focus on molecular dynamics. JAX MD includes a number of statistical physics simulation environments as well as interaction potentials and neural networks that can be integrated into these environments without writing any additional code. Since the simulations themselves are differentiable functions, entire trajectories can be differentiated to perform meta-optimization. These features are built on primitive operations, such as spatial partitioning, that allow simulations to scale to hundreds-of-thousands of particles on a single GPU. These primitives are flexible enough that they can be used to scale up workloads outside of molecular dynamics. We present several examples that highlight the features of JAX MD including: integration of graph neural networks into traditional simulations, meta-optimization through minimization of particle packings, and a multi-agent flocking simulation. JAX MD is available at www.github.com/google/jax-md.

[An implicit function learning approach for parametric modal regression](#)

- Yangchen Pan, Ehsan Imani, Amir-massoud Farahmand, Martha White
- abstract@[open-review](#): For multi-valued functions---such as when the conditional distribution on targets given the inputs is multi-modal---standard regression approaches are not always desirable because they provide the conditional mean. Modal regression algorithms address this issue by instead finding the conditional mode(s). Most, however, are nonparametric approaches and so can be difficult to scale. Further, parametric approximators, like neural networks, facilitate learning complex relationships between inputs and targets. In this work, we propose a parametric modal regression algorithm. We use the implicit function theorem to develop an objective, for learning a joint function over inputs and targets. We empirically demonstrate on several synthetic problems that our method (i) can learn multi-valued functions and produce the conditional modes, (ii) scales well to high-dimensional inputs, and (iii) can even be more effective for certain uni-modal problems, particularly for high-frequency functions. We demonstrate that our method is competitive in a real-world modal regression problem and two regular regression datasets.

[SDF-SRN: Learning Signed Distance 3D Object Reconstruction from Static Images](#)

- Chen-Hsuan Lin, Chaoyang Wang, Simon Lucey
- abstract@[open-review](#): Dense 3D object reconstruction from a single image has recently witnessed remarkable advances, but supervising neural networks with ground-truth 3D shapes is impractical due to the laborious process of creating paired image-shape datasets. Recent efforts have turned to learning 3D reconstruction without 3D supervision from RGB images with annotated 2D silhouettes, dramatically reducing the cost and effort of annotation. These techniques, however, remain impractical as they still require multi-view annotations of the same object instance during training. As a result, most experimental efforts to date have been limited to synthetic datasets. In this paper, we address this issue and propose SDF-SRN, an approach that requires only a single view of objects at training time, offering greater utility for real-world scenarios. SDF-SRN learns implicit 3D shape representations to handle arbitrary shape topologies that may exist in the datasets. To this end, we derive a novel differentiable rendering formulation for learning signed distance functions (SDF) from 2D silhouettes. Our method outperforms the state of the art under challenging single-view supervision settings on both synthetic and real-world datasets.

[Coresets for Robust Training of Deep Neural Networks against Noisy Labels](#)

- Baharan Mirzasoleiman, Kaidi Cao, Jure Leskovec
- abstract@[open-review](#): Modern neural networks have the capacity to overfit noisy labels frequently found in real-world datasets. Although great progress has been made, existing techniques are very limited in providing theoretical guarantees for the performance of the neural networks trained with noisy labels. To tackle this challenge, we propose a novel approach with strong theoretical guarantees for robust training of neural networks trained with noisy labels. The key idea behind our method is to select subsets of clean data points that provide an approximately low-rank Jacobian matrix. We then prove that gradient descent applied to the subsets cannot overfit the noisy labels, without regularization or early stopping. Our extensive experiments corroborate our theory and demonstrate that deep networks trained on our subsets achieve a significantly superior performance, e.g., 7% increase in accuracy on mini Webvision with 50% noisy labels, compared to state-of-the art.

[Adapting to Misspecification in Contextual Bandits](#)

- Dylan J. Foster, Claudio Gentile, Mehryar Mohri, Julian Zimmert
- abstract@[open-review](#): A major research direction in contextual bandits is to develop algorithms that are computationally efficient, yet support flexible, general-purpose function approximation. Algorithms based on modeling rewards have shown strong empirical performance, yet typically require a well-specified model, and can fail when this assumption does not hold. Can we design algorithms that are efficient and flexible, yet degrade gracefully in the face of model misspecification? We introduce a new family of oracle-efficient algorithms for $\backslash varepsilon$ -misspecified contextual bandits that adapt to unknown model misspecification---both for finite and infinite action settings. Given access to an $\backslash emph{online oracle}$ for square loss regression, our algorithm attains optimal regret and---in particular---optimal dependence on the misspecification level, with $\backslash emph{no prior knowledge}$. Specializing to linear contextual bandits with infinite actions in $\$d\$$ dimensions, we obtain the first algorithm that achieves the optimal $\backslash bigoh(d\sqrt{T} + \backslash varepsilon\sqrt{d}T)$ regret bound for unknown $\backslash varepsilon$.

On a conceptual level, our results are enabled by a new optimization-based perspective on the regression oracle reduction framework of Foster and Rakhlin (2020), which we believe will be useful more broadly.

[Convergence of Meta-Learning with Task-Specific Adaptation over Partial Parameters](#)

- Kaiyi Ji, Jason D. Lee, Yingbin Liang, H. Vincent Poor
- abstract@[open-review](#): Although model-agnostic meta-learning (MAML) is a very successful algorithm in meta-learning practice, it can have high computational cost because it updates all model parameters over both the inner loop of task-specific adaptation and the outer-loop of meta initialization training. A more efficient algorithm ANIL (which refers to almost no inner loop) was proposed recently by Raghu et al. 2019, which adapts only a small subset of parameters in the inner loop and thus has substantially less computational cost than MAML as demonstrated by extensive experiments. However, the theoretical convergence of ANIL has not been studied yet. In this paper, we characterize the convergence rate and the computational complexity for ANIL under two representative inner-loop loss geometries, i.e., strongly-convexity and nonconvexity. Our results show that such a geometric property can significantly affect the overall convergence performance of ANIL. For example, ANIL achieves a faster convergence rate for a strongly-convex inner-loop loss as the number $\$N\$$ of inner-loop gradient descent steps increases, but a slower convergence rate for a nonconvex inner-loop loss as $\$N\$$ increases. Moreover, our complexity analysis provides a theoretical quantification on the improved efficiency of ANIL over MAML. The experiments on standard few-shot meta-learning benchmarks validate our theoretical findings.

[MetaPerturb: Transferable Regularizer for Heterogeneous Tasks and Architectures](#)

- Jeong Un Ryu, JaeWoong Shin, Hae Beom Lee, Sung Ju Hwang
- abstract@[open-review](#): Regularization and transfer learning are two popular techniques to enhance model generalization on unseen data, which is a fundamental problem of machine learning. Regularization techniques are versatile, as they are task- and architecture-agnostic, but they do not exploit a large amount of data available. Transfer learning methods learn to transfer knowledge from one domain to another, but may not generalize across tasks and architectures, and may introduce new training cost for adapting to the target task. To bridge the gap between the two, we propose a transferable perturbation, MetaPerturb, which is meta-learned to improve generalization performance on unseen data. MetaPerturb is implemented as a set-based lightweight network that is agnostic to the size and the order of the input, which is shared across the layers. Then, we propose a meta-learning framework, to jointly train the perturbation function over heterogeneous tasks in parallel. As MetaPerturb is a set-function trained over diverse distributions across layers and tasks, it can generalize to heterogeneous tasks and architectures. We validate the efficacy and generality of MetaPerturb trained on a specific source domain and architecture, by applying it to the training of diverse neural architectures on heterogeneous target datasets against various regularizers and fine-tuning. The results show that the networks trained with MetaPerturb significantly outperform the baselines on most of the tasks and architectures, with a negligible increase in the parameter size and no hyperparameters to tune.

[Learning to solve TV regularised problems with unrolled algorithms](#)

- Hamza Cherkaoui, Jeremias Sulam, Thomas Moreau

- abstract@[open-review](#): Total Variation (TV) is a popular regularization strategy that promotes piece-wise constant signals by constraining the ℓ_1 -norm of the first order derivative of the estimated signal. The resulting optimization problem is usually solved using iterative algorithms such as proximal gradient descent, primal-dual algorithms or ADMM. However, such methods can require a very large number of iterations to converge to a suitable solution. In this paper, we accelerate such iterative algorithms by unfolding proximal gradient descent solvers in order to learn their parameters for 1D TV regularized problems. While this could be done using the synthesis formulation, we demonstrate that this leads to slower performances. The main difficulty in applying such methods in the analysis formulation lies in proposing a way to compute the derivatives through the proximal operator. As our main contribution, we develop and characterize two approaches to do so, describe their benefits and limitations, and discuss the regime where they can actually improve over iterative procedures. We validate those findings with experiments on synthetic and real data.

Object-Centric Learning with Slot Attention

- Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, Thomas Kipf
- abstract@[open-review](#): Learning object-centric representations of complex scenes is a promising step towards enabling efficient abstract reasoning from low-level perceptual features. Yet, most deep learning approaches learn distributed representations that do not capture the compositional properties of natural scenes. In this paper, we present the Slot Attention module, an architectural component that interfaces with perceptual representations such as the output of a convolutional neural network and produces a set of task-dependent abstract representations which we call slots. These slots are exchangeable and can bind to any object in the input by specializing through a competitive procedure over multiple rounds of attention. We empirically demonstrate that Slot Attention can extract object-centric representations that enable generalization to unseen compositions when trained on unsupervised object discovery and supervised property prediction tasks.

Improving robustness against common corruptions by covariate shift adaptation

- Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, Matthias Bethge
- abstract@[open-review](#): Today's state-of-the-art machine vision models are vulnerable to image corruptions like blurring or compression artefacts, limiting their performance in many real-world applications. We here argue that popular benchmarks to measure model robustness against common corruptions (like ImageNet-C) underestimate model robustness in many (but not all) application scenarios. The key insight is that in many scenarios, multiple unlabeled examples of the corruptions are available and can be used for unsupervised online adaptation. Replacing the activation statistics estimated by batch normalization on the training set with the statistics of the corrupted images consistently improves the robustness across 25 different popular computer vision models. Using the corrected statistics, ResNet-50 reaches 62.2% mCE on ImageNet-C compared to 76.7% without adaptation. With the more robust DeepAugment+AugMix model, we improve the state of the art achieved by a ResNet50 model up to date from 53.6% mCE to 45.4% mCE. Even adapting to a single sample improves robustness for the ResNet-50 and AugMix models, and 32 samples are sufficient to improve the current state of the art for a ResNet-50 architecture. We argue that results with adapted statistics should be included whenever reporting scores in corruption benchmarks and other out-of-distribution generalization settings.

Deep Smoothing of the Implied Volatility Surface

- Damien Ackerer, Natasa Tagasovska, Thibault Vatter
- abstract@[open-review](#): We present a neural network (NN) approach to fit and predict implied volatility surfaces (IVSs). Atypically to standard NN applications, financial industry practitioners use such models equally to replicate market prices and to value other financial instruments. In other words, low training losses are as important as generalization capabilities. Importantly, IVS models need to generate realistic arbitrage-free option prices, meaning that no portfolio can lead to risk-free profits. We propose an approach guaranteeing the absence of arbitrage opportunities by penalizing the loss using soft constraints. Furthermore, our method can be combined with standard IVS models in quantitative finance, thus providing a NN-based correction when such models fail at replicating observed market prices. This lets practitioners use our approach as a plug-in on top of classical methods. Empirical results show that this approach is particularly useful when only sparse or erroneous data are available. We also quantify the uncertainty of the model predictions in regions with few or no observations. We further explore how deeper NNs improve over shallower ones, as well as other properties of the network architecture. We benchmark our method against standard IVS models. By evaluating our method on both training sets, and testing sets, namely, we highlight both their capacity to reproduce observed prices and predict new ones.

Probabilistic Inference with Algebraic Constraints: Theoretical Limits and Practical Approximations

- Zhe Zeng, Paolo Morettin, Fanqi Yan, Antonio Vergari, Guy Van den Broeck
- abstract@[open-review](#): Weighted model integration (WMI) is a framework to perform advanced probabilistic inference on hybrid domains, i.e., on distributions over mixed continuous-discrete random variables and in presence of complex logical and arithmetic constraints. In this work, we advance the WMI framework on both the theoretical and algorithmic side. First, we exactly trace the boundaries of tractability for WMI inference by proving that to be amenable to exact and efficient inference a WMI problem has to possess a tree-shaped structure with logarithmic diameter. While this result deepens our theoretical understanding of WMI it hinders the practical applicability of exact WMI solvers to real-world problems. To overcome this, we propose the first approximate WMI solver that does not resort to sampling, but performs exact inference on one approximate models. Our solution performs message passing in a relaxed problem structure iteratively to recover certain lost dependencies and, as our experiments suggest, is competitive with other SOTA WMI solvers.

Provable Online CP/PARAFAC Decomposition of a Structured Tensor via Dictionary Learning

- Sirisha Rambhatla, Xingguo Li, Jarvis Haupt
- abstract@[open-review](#): We consider the problem of factorizing a structured 3-way tensor into its constituent Canonical Polyadic (CP) factors. This decomposition, which can be viewed as a generalization of singular value decomposition (SVD) for tensors, reveals how the tensor dimensions (features) interact with each other. However, since the factors are a priori unknown, the corresponding optimization problems are inherently non-convex. The existing guaranteed algorithms which handle this non-convexity incur an irreducible error (bias), and only apply to cases where all factors have the same structure. To this end, we develop a provable algorithm for online structured tensor factorization, wherein one of the factors obeys some incoherence conditions, and the others are sparse. Specifically we show that, under some relatively mild conditions on initialization, rank, and sparsity, our algorithm recovers the factors exactly (up to scaling and permutation) at a linear rate. Complementary to our theoretical results, our synthetic and real-world data evaluations showcase superior performance compared to related techniques.

Look-ahead Meta Learning for Continual Learning

- Gunshi Gupta, Karmesh Yadav, Liam Paull
- abstract@[open-review](#): The continual learning problem involves training models with limited capacity to perform well on a set of an unknown number of sequentially arriving tasks. While meta-learning shows great potential for reducing interference between old and new tasks, the current training procedures tend to be either slow or offline, and sensitive to many hyper-parameters. In this work, we propose Look-ahead MAML (La-MAML), a fast optimisation-based meta-learning algorithm for online-continual learning, aided by a small episodic memory. By incorporating the modulation of per-parameter learning rates in our meta-learning update, our approach also allows us to draw connections to and exploit prior work on hypergradients and meta-descent. This provides a more flexible and efficient way to mitigate catastrophic forgetting compared to conventional prior-based methods. La-MAML achieves

performance superior to other replay-based, prior-based and meta-learning based approaches for continual learning on real-world visual classification benchmarks.

[A polynomial-time algorithm for learning nonparametric causal graphs](#)

- Ming Gao, Yi Ding, Bryon Aragam
- abstract@[open-review](#): We establish finite-sample guarantees for a polynomial-time algorithm for learning a nonlinear, nonparametric directed acyclic graphical (DAG) model from data. The analysis is model-free and does not assume linearity, additivity, independent noise, or faithfulness. Instead, we impose a condition on the residual variances that is closely related to previous work on linear models with equal variances. Compared to an optimal algorithm with oracle knowledge of the variable ordering, the additional cost of the algorithm is linear in the dimension d and the number of samples n . Finally, we compare the proposed algorithm to existing approaches in a simulation study.

[Sparse Learning with CART](#)

- Jason Klusowski
- abstract@[open-review](#): Decision trees with binary splits are popularly constructed using Classification and Regression Trees (CART) methodology. For regression models, this approach recursively divides the data into two near-homogenous daughter nodes according to a split point that maximizes the reduction in sum of squares error (the impurity) along a particular variable. This paper aims to study the statistical properties of regression trees constructed with CART. In doing so, we find that the training error is governed by the Pearson correlation between the optimal decision stump and response data in each node, which we bound by constructing a prior distribution on the split points and solving a nonlinear optimization problem. We leverage this connection between the training error and Pearson correlation to show that CART with cost-complexity pruning achieves an optimal complexity/goodness-of-fit tradeoff when the depth scales with the logarithm of the sample size. Data dependent quantities, which adapt to the dimensionality and latent structure of the regression model, are seen to govern the rates of convergence of the prediction error.

[Proximal Mapping for Deep Regularization](#)

- Mao Li, Yingyi Ma, Xinhua Zhang
- abstract@[open-review](#): Underpinning the success of deep learning is effective regularizations that allow a variety of priors in data to be modeled. For example, robustness to adversarial perturbations, and correlations between multiple modalities. However, most regularizers are specified in terms of hidden layer outputs, which are not themselves optimization variables. In contrast to prevalent methods that optimize them indirectly through model weights, we propose inserting proximal mapping as a new layer to the deep network, which directly and explicitly produces well regularized hidden layer outputs. The resulting technique is shown well connected to kernel warping and dropout, and novel algorithms were developed for robust temporal learning and multiview modeling, both outperforming state-of-the-art methods.

[Identifying Causal-Effect Inference Failure with Uncertainty-Aware Models](#)

- Andrew Jesson, Sören Mindermann, Uri Shalit, Yarin Gal
- abstract@[open-review](#): Recommending the best course of action for an individual is a major application of individual-level causal effect estimation. This application is often needed in safety-critical domains such as healthcare, where estimating and communicating uncertainty to decision-makers is crucial. We introduce a practical approach for integrating uncertainty estimation into a class of state-of-the-art neural network methods used for individual-level causal estimates. We show that our methods enable us to deal gracefully with situations of "no-overlap", common in high-dimensional data, where standard applications of causal effect approaches fail. Further, our methods allow us to handle covariate shift, where the train and test distributions differ, common when systems are deployed in practice. We show that when such a covariate shift occurs, correctly modeling uncertainty can keep us from giving overconfident and potentially harmful recommendations. We demonstrate our methodology with a range of state-of-the-art models. Under both covariate shift and lack of overlap, our uncertainty-equipped methods can alert decision makers when predictions are not to be trusted while outperforming standard methods that use the propensity score to identify lack of overlap.

[Hierarchical Granularity Transfer Learning](#)

- Shaobo Min, Hongtao Xie, Hantao Yao, Xuran Deng, Zheng-Jun Zha, Yongdong Zhang
- abstract@[open-review](#): In the real world, object categories usually have a hierarchical granularity tree. Nowadays, most researchers focus on recognizing categories in a specific granularity, e.g., basic-level or sub(ordinate)-level. Compared with basic-level categories, the sub-level categories provide more valuable information, but its training annotations are harder to acquire. Therefore, an attractive problem is how to transfer the knowledge learned from basic-level annotations to sub-level recognition. In this paper, we introduce a new task, named Hierarchical Granularity Transfer Learning (HGTL), to recognize sub-level categories with basic-level annotations and semantic descriptions for hierarchical categories. Different from other recognition tasks, HGTL has a serious granularity gap, i.e., the two granularities share an image space but have different category domains, which impede the knowledge transfer. To this end, we propose a novel Bi-granularity Semantic Preserving Network (BigSPN) to bridge the granularity gap for robust knowledge transfer. Explicitly, BigSPN constructs specific visual encoders for different granularities, which are aligned with a shared semantic interpreter via a novel subordinate entropy loss. Experiments on three benchmarks with hierarchical granularities show that BigSPN is an effective framework for Hierarchical Granularity Transfer Learning.

[Deep active inference agents using Monte-Carlo methods](#)

- Zafeirios Fountas, Noor Sajid, Pedro Mediano, Karl Friston
- abstract@[open-review](#): Active inference is a Bayesian framework for understanding biological intelligence. The underlying theory brings together perception and action under one single imperative: minimizing free energy. However, despite its theoretical utility in explaining intelligence, computational implementations have been restricted to low-dimensional and idealized situations. In this paper, we present a neural architecture for building deep active inference agents operating in complex, continuous state-spaces using multiple forms of Monte-Carlo (MC) sampling. For this, we introduce a number of techniques, novel to active inference. These include: i) selecting free-energy-optimal policies via MC tree search, ii) approximating this optimal policy distribution via a feed-forward 'habitual' network, iii) predicting future parameter belief updates using MC dropouts and, finally, iv) optimizing state transition precision (a high-end form of attention). Our approach enables agents to learn environmental dynamics efficiently, while maintaining task performance, in relation to reward-based counterparts. We illustrate this in a new toy environment, based on the dSprites data-set, and demonstrate that active inference agents automatically create disentangled representations that are apt for modeling state transitions. In a more complex Animal-AI environment, our agents (using the same neural architecture) are able to simulate future state transitions and actions (i.e., plan), to evince reward-directed navigation - despite temporary suspension of visual input. These results show that deep active inference - equipped with MC methods - provides a flexible framework to develop biologically-inspired intelligent agents, with applications in both machine learning and cognitive science.

[Consistent Estimation of Identifiable Nonparametric Mixture Models from Grouped Observations](#)

- Alexander Ritchie, Robert A. Vandermeulen, Clayton Scott

- abstract@[open-review](#): Recent research has established sufficient conditions for finite mixture models to be identifiable from grouped observations. These conditions allow the mixture components to be nonparametric and have substantial (or even total) overlap. This work proposes an algorithm that consistently estimates any identifiable mixture model from grouped observations. Our analysis leverages an oracle inequality for weighted kernel density estimators of the distribution on groups, together with a general result showing that consistent estimation of the distribution on groups implies consistent estimation of mixture components. A practical implementation is provided for paired observations, and the approach is shown to outperform existing methods, especially when mixture components overlap significantly.

[Manifold structure in graph embeddings](#)

- Patrick Rubin-Delanchy
- abstract@[open-review](#): Statistical analysis of a graph often starts with embedding, the process of representing its nodes as points in space. How to choose the embedding dimension is a nuanced decision in practice, but in theory a notion of true dimension is often available. In spectral embedding, this dimension may be very high. However, this paper shows that existing random graph models, including graphon and other latent position models, predict the data should live near a much lower-dimensional set. One may therefore circumvent the curse of dimensionality by employing methods which exploit hidden manifold structure.

[Adaptive Learned Bloom Filter \(Ada-BF\): Efficient Utilization of the Classifier with Application to Real-Time Information Filtering on the Web](#)

- Zhenwei Dai, Anshumali Shrivastava
- abstract@[open-review](#): Recent work suggests improving the performance of Bloom filter by incorporating a machine learning model as a binary classifier. However, such learned Bloom filter does not take full advantage of the predicted probability scores. We propose new algorithms that generalize the learned Bloom filter by using the complete spectrum of the score regions. We prove our algorithms have lower false positive rate (FPR) and memory usage compared with the existing approaches to learned Bloom filter. We also demonstrate the improved performance of our algorithms on real-world information filtering tasks over the web.

[MCUNet: Tiny Deep Learning on IoT Devices](#)

- Ji Lin, Wei-Ming Chen, Yujun Lin, John Cohn, Chuang Gan, Song Han
- abstract@[open-review](#): Machine learning on tiny IoT devices based on microcontroller units (MCU) is appealing but challenging: the memory of microcontrollers is 2-3 orders of magnitude smaller even than mobile phones. We propose MCUNet, a framework that jointly designs the efficient neural architecture (TinyNAS) and the lightweight inference engine (TinyEngine), enabling ImageNet-scale inference on microcontrollers. TinyNAS adopts a two-stage neural architecture search approach that first optimizes the search space to fit the resource constraints, then specializes the network architecture in the optimized search space. TinyNAS can automatically handle diverse constraints (i.e. device, latency, energy, memory) under low search costs. TinyNAS is co-designed with TinyEngine, a memory-efficient inference library to expand the search space and fit a larger model. TinyEngine adapts the memory scheduling according to the overall network topology rather than layer-wise optimization, reducing the memory usage by 3.4 \AA , and accelerating the inference by 1.7-3.3 \AA — compared to TF-Lite Micro [3] and CMSIS-NN [28]. MCUNet is the first to achieve >70% ImageNet top1 accuracy on an off-the-shelf commercial microcontroller, using 3.5 \AA — less SRAM and 5.7 \AA — less Flash compared to quantized MobileNetV2 and ResNet-18. On visual&audio wake words tasks, MCUNet achieves state-of-the-art accuracy and runs 2.4-3.4 \AA — faster than MobileNetV2 and ProxylessNAS-based solutions with 3.7-4.1 \AA — smaller peak SRAM. Our study suggests that the era of always-on tiny machine learning on IoT devices has arrived.

[In search of robust measures of generalization](#)

- Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, Daniel M. Roy
- abstract@[open-review](#): One of the principal scientific challenges in deep learning is explaining generalization, i.e., why the particular way the community now trains networks to achieve small training error also leads to small error on held-out data from the same population. It is widely appreciated that some worst-case theories — such as those based on the VC dimension of the class of predictors induced by modern neural network architectures — are unable to explain empirical performance. A large volume of work aims to close this gap, primarily by developing bounds on generalization error, optimization error, and excess risk. When evaluated empirically, however, most of these bounds are numerically vacuous. Focusing on generalization bounds, this work addresses the question of how to evaluate such bounds empirically. Jiang et al. (2020) recently described a large-scale empirical study aimed at uncovering potential causal relationships between bounds/measures and generalization. Building on their study, we highlight where their proposed methods can obscure failures and successes of generalization measures in explaining generalization. We argue that generalization measures should instead be evaluated within the framework of distributional robustness.

[Task-agnostic Exploration in Reinforcement Learning](#)

- Xuezhou Zhang, Yuzhe Ma, Adish Singla
- abstract@[open-review](#): Efficient exploration is one of the main challenges in reinforcement learning (RL). Most existing sample-efficient algorithms assume the existence of a single reward function during exploration. In many practical scenarios, however, there is not a single underlying reward function to guide the exploration, for instance, when an agent needs to learn many skills simultaneously, or multiple conflicting objectives need to be balanced. To address these challenges, we propose the \textit{task-agnostic RL} framework: In the exploration phase, the agent first collects trajectories by exploring the MDP without the guidance of a reward function. After exploration, it aims at finding near-optimal policies for $\$N\$$ tasks, given the collected trajectories augmented with \textit{sampled rewards} for each task. We present an efficient task-agnostic RL algorithm, \textit{UCBZero}, that finds $\$epsilon$ -optimal policies for $\$N\$$ arbitrary tasks after at most $\tilde{O}(\log(N)H^5SA/\epsilon^2)$ exploration episodes. We also provide an $\Omega(\log(N)H^2SA/\epsilon^2)$ lower bound, showing that the \log dependency on $\$N\$$ is unavoidable. Furthermore, we provide an $\$N\$$ -independent sample complexity bound of \textit{UCBZero} in the statistically easier setting when the ground truth reward functions are known.

[Multi-task Additive Models for Robust Estimation and Automatic Structure Discovery](#)

- Yingjie Wang, Hong Chen, Feng Zheng, Chen Xu, Teliang Gong, Yanhong Chen
- abstract@[open-review](#): Additive models have attracted much attention for high-dimensional regression estimation and variable selection. However, the existing models are usually limited to the single-task learning framework under the mean squared error (MSE) criterion, where the utilization of variable structure depends heavily on prior knowledge among variables. For high-dimensional observations in real environment, e.g., Coronal Mass Ejections (CMEs) data, the learning performance of previous methods may be degraded seriously due to the complex non-Gaussian noise and the insufficiency of prior knowledge on variable structure. To tackle this problem, we propose a new class of additive models, called Multi-task Additive Models (MAM), by integrating the mode-induced metric, the structure-based regularizer, and additive hypothesis spaces into a bilevel optimization framework. Our approach does not require any prior knowledge of variable structure and suits for high-dimensional data with complex noise, e.g., skewed noise, heavy-tailed noise, and outliers. A smooth iterative optimization algorithm with convergence guarantees is provided to implement MAM efficiently. Experiments on simulations and the CMEs analysis demonstrate the competitive performance of our approach for robust estimation and automatic structure discovery.

Provably Efficient Reward-Agnostic Navigation with Linear Value Iteration

- Andrea Zanette, Alessandro Lazaric, Mykel J. Kochenderfer, Emma Brunskill
- abstract@[open-review](#): There has been growing progress on theoretical analyses for provably efficient learning in MDPs with linear function approximation, but much of the existing work has made strong assumptions to enable exploration by conventional exploration frameworks. Typically these assumptions are stronger than what is needed to find good solutions in the batch setting. In this work, we show how under a more standard notion of low inherent Bellman error, typically employed in least-square value iteration-style algorithms, we can provide strong PAC guarantees on learning a near optimal value function provided that the linear space is sufficiently ``explorable''. We present a computationally tractable algorithm for the reward-free setting and show how it can be used to learn a near optimal policy for any (linear) reward function, which is revealed only once learning has completed. If this reward function is also estimated from the samples gathered during pure exploration, our results also provide same-order PAC guarantees on the performance of the resulting policy for this setting.

Softmax Deep Double Deterministic Policy Gradients

- Ling Pan, Qingpeng Cai, Longbo Huang
- abstract@[open-review](#): A widely-used actor-critic reinforcement learning algorithm for continuous control, Deep Deterministic Policy Gradients (DDPG), suffers from the overestimation problem, which can negatively affect the performance. Although the state-of-the-art Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm mitigates the overestimation issue, it can lead to a large underestimation bias. In this paper, we propose to use the Boltzmann softmax operator for value function estimation in continuous control. We first theoretically analyze the softmax operator in continuous action space. Then, we uncover an important property of the softmax operator in actor-critic algorithms, i.e., it helps to smooth the optimization landscape, which sheds new light on the benefits of the operator. We also design two new algorithms, Softmax Deep Deterministic Policy Gradients (SD2) and Softmax Deep Double Deterministic Policy Gradients (SD3), by building the softmax operator upon single and double estimators, which can effectively improve the overestimation and underestimation bias. We conduct extensive experiments on challenging continuous control tasks, and results show that SD3 outperforms state-of-the-art methods.

Online Decision Based Visual Tracking via Reinforcement Learning

- ke Song, Wei Zhang, Ran Song, Yibin Li
- abstract@[open-review](#): A deep visual tracker is typically based on either object detection or template matching while each of them is only suitable for a particular group of scenes. It is straightforward to consider fusing them together to pursue more reliable tracking. However, this is not wise as they follow different tracking principles. Unlike previous fusion-based methods, we propose a novel ensemble framework, named DTNet, with an online decision mechanism for visual tracking based on hierarchical reinforcement learning. The decision mechanism substantiates an intelligent switching strategy where the detection and the template trackers have to compete with each other to conduct tracking within different scenes that they are adept in. Besides, we present a novel detection tracker which avoids the common issue of incorrect proposal. Extensive results show that our DTNet achieves state-of-the-art tracking performance as well as good balance between accuracy and efficiency. The project website is available at <https://vslab.github.io/DTNet/>.

Efficient Marginalization of Discrete and Structured Latent Variables via Sparsity

- Gonçalo Correia, Vlad Niculae, Wilker Aziz, André Martins
- abstract@[open-review](#): Training neural network models with discrete (categorical or structured) latent variables can be computationally challenging, due to the need for marginalization over large or combinatorial sets. To circumvent this issue, one typically resorts to sampling-based approximations of the true marginal, requiring noisy gradient estimators (e.g., score function estimator) or continuous relaxations with lower-variance reparameterized gradients (e.g., Gumbel-Softmax). In this paper, we propose a new training strategy which replaces these estimators by an exact yet efficient marginalization. To achieve this, we parameterize discrete distributions over latent assignments using differentiable sparse mappings: sparsemax and its structured counterparts. In effect, the support of these distributions is greatly reduced, which enables efficient marginalization. We report successful results in three tasks covering a range of latent variable modeling applications: a semisupervised deep generative model, a latent communication game, and a generative model with a bit-vector latent representation. In all cases, we obtain good performance while still achieving the practicality of sampling-based approximations.

DeepI2I: Enabling Deep Hierarchical Image-to-Image Translation by Transferring from GANs

- yaxing wang, Lu Yu, Joost van de Weijer
- abstract@[open-review](#): Image-to-image translation has recently achieved remarkable results. But despite current success, it suffers from inferior performance when translations between classes require large shape changes. We attribute this to the high-resolution bottlenecks which are used by current state-of-the-art image-to-image methods. Therefore, in this work, we propose a novel deep hierarchical Image-to-Image Translation method, called DeepI2I. We learn a model by leveraging hierarchical features: (a) structural information contained in the bottom layers and (b) semantic information extracted from the top layers. To enable the training of deep I2I models on small datasets, we propose a novel transfer learning method, that transfers knowledge from pre-trained GANs. Specifically, we leverage the discriminator of a pre-trained GANs (i.e. BigGAN or StyleGAN) to initialize both the encoder and the discriminator and the pre-trained generator to initialize the generator of our model. Applying knowledge transfer leads to an alignment problem between the encoder and generator. We introduce an adaptor network to address this. On many-class image-to-image translation on three datasets (Animal faces, Birds, and Foods) we decrease mFID by at least 35% when compared to the state-of-the-art. Furthermore, we qualitatively and quantitatively demonstrate that transfer learning significantly improves the performance of I2I systems, especially for small datasets. Finally, we are the first to perform I2I translations for domains with over 100 classes.

Distributional Robustness with IPMs and links to Regularization and GANs

- Hisham Husain
- abstract@[open-review](#): Robustness to adversarial attacks is an important concern due to the fragility of deep neural networks to small perturbations, and has received an abundance of attention in recent years. Distributional Robust Optimization (DRO), a particularly promising way of addressing this challenge, studies robustness via divergence-based uncertainty sets and has provided valuable insights into robustification strategies such as regularisation. In the context of machine learning, majority of existing results have chosen $\$f\$$ -divergences, Wasserstein distances and more recently, the Maximum Mean Discrepancy (MMD) to construct uncertainty sets. We extend this line of work for the purposes of understanding robustness via regularization by studying uncertainty sets constructed with Integral Probability Metrics (IPMs) - a large family of divergences including the MMD, Total Variation and Wasserstein distances. Our main result shows that DRO under \textit{any} choice of IPM corresponds to a family of regularization penalties, which recover and improve upon existing results in the setting of MMD and Wasserstein distances. Due to the generality of our result, we show that other choices of IPMs correspond to other commonly used penalties in machine learning. Furthermore, we extend our results to shed light on adversarial generative modelling via $\$f\$$ -GANs, constituting the first study of distributional robustness for the $\$f\$$ -GAN objective. Our results unveil the inductive properties of the discriminator set with regards to robustness, allowing us to give positive comments for a number of existing penalty-based GAN methods such as Wasserstein-, MMD- and Sobolev-GANs. In summary, our results intimately link GANs to distributional robustness, extend previous results on DRO and contribute to our understanding of the link between regularization and robustness at large.

A shooting formulation of deep learning

- FranÃ§ois-Xavier Vialard, Roland Kwitt, Susan Wei, Marc Niethammer
- abstract@[open-review](#): A residual network may be regarded as a discretization of an ordinary differential equation (ODE) which, in the limit of time discretization, defines a continuous-depth network. Although important steps have been taken to realize the advantages of such continuous formulations, most current techniques assume identical layers. Indeed, existing works throw into relief the myriad difficulties of learning an infinite-dimensional parameter in a continuous-depth neural network. To this end, we introduce a shooting formulation which shifts the perspective from parameterizing a network layer-by-layer to parameterizing over optimal networks described only by a set of initial conditions. For scalability, we propose a novel particle-ensemble parameterization which fully specifies the optimal weight trajectory of the continuous-depth neural network. Our experiments show that our particle-ensemble shooting formulation can achieve competitive performance. Finally, though the current work is inspired by continuous-depth neural networks, the particle-ensemble shooting formulation also applies to discrete-time networks and may lead to a new fertile area of research in deep learning parameterization.

[CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances](#)

- Jihoon Tack, Sangwoo Mo, Jongheon Jeong, Jinwoo Shin
- abstract@[open-review](#): Novelty detection, i.e., identifying whether a given sample is drawn from outside the training distribution, is essential for reliable machine learning. To this end, there have been many attempts at learning a representation well-suited for novelty detection and designing a score based on such representation. In this paper, we propose a simple, yet effective method named contrasting shifted instances (CSI), inspired by the recent success on contrastive learning of visual representations. Specifically, in addition to contrasting a given sample with other instances as in conventional contrastive learning methods, our training scheme contrasts the sample with distributionally-shifted augmentations of itself. Based on this, we propose a new detection score that is specific to the proposed training scheme. Our experiments demonstrate the superiority of our method under various novelty detection scenarios, including unlabeled one-class, unlabeled multi-class and labeled multi-class settings, with various image benchmark datasets. Code and pre-trained models are available at <https://github.com/alinlab/CSI>.

[Learning Implicit Credit Assignment for Cooperative Multi-Agent Reinforcement Learning](#)

- Meng Zhou, Ziyu Liu, Pengwei Sui, Yixuan Li, Yuk Ying Chung
- abstract@[open-review](#): We present a multi-agent actor-critic method that aims to implicitly address the credit assignment problem under fully cooperative settings. Our key motivation is that credit assignment among agents may not require an explicit formulation as long as (1) the policy gradients derived from a centralized critic carry sufficient information for the decentralized agents to maximize their joint action value through optimal cooperation and (2) a sustained level of exploration is enforced throughout training. Under the centralized training with decentralized execution (CTDE) paradigm, we achieve the former by formulating the centralized critic as a hypernetwork such that a latent state representation is integrated into the policy gradients through its multiplicative association with the stochastic policies; to achieve the latter, we derive a simple technique called adaptive entropy regularization where magnitudes of the entropy gradients are dynamically rescaled based on the current policy stochasticity to encourage consistent levels of exploration. Our algorithm, referred to as LICA, is evaluated on several benchmarks including the multi-agent particle environments and a set of challenging StarCraft II micromanagement tasks, and we show that LICA significantly outperforms previous methods.

[MATE: Plugging in Model Awareness to Task Embedding for Meta Learning](#)

- Xiaohan Chen, Zhangyang Wang, Siyu Tang, Krikamol Muandet
- abstract@[open-review](#): Meta-learning improves generalization of machine learning models when faced with previously unseen tasks by leveraging experiences from different, yet related prior tasks. To allow for better generalization, we propose a novel task representation called model-aware task embedding (MATE) that incorporates not only the data distributions of different tasks, but also the complexity of the tasks through the models used. The task complexity is taken into account by a novel variant of kernel mean embedding, combined with an instance-adaptive attention mechanism inspired by an SVM-based feature selection algorithm. Together with conditioning layers in deep neural networks, MATE can be easily incorporated into existing meta learners as a plug-and-play module. While MATE is widely applicable to general tasks where the concept of task/environment is involved, we demonstrate its effectiveness in few-shot learning by improving a state-of-the-art model consistently on two benchmarks. Source codes for this paper are available at <https://github.com/VITA-Group/MATE>.

[Restless-UCB, an Efficient and Low-complexity Algorithm for Online Restless Bandits](#)

- Siwei Wang, Longbo Huang, John C. S. Lui
- abstract@[open-review](#): We study the online restless bandit problem, where the state of each arm evolves according to a Markov chain, and the reward of pulling an arm depends on both the pulled arm and the current state of the corresponding Markov chain. In this paper, we propose Restless-UCB, a learning policy that follows the explore-then-commit framework. In Restless-UCB, we present a novel method to construct offline instances, which only requires $\$O(N)$ time-complexity (N is the number of arms) and is exponentially better than the complexity of existing learning policy. We also prove that Restless-UCB achieves a regret upper bound of $\tilde{O}((N+M^3)T^{2/3})$, where M is the Markov chain state space size and T is the time horizon. Compared to existing algorithms, our result eliminates the exponential factor (in M, N) in the regret upper bound, due to a novel exploitation of the sparsity in transitions in general restless bandit problems. As a result, our analysis technique can also be adopted to tighten the regret bounds of existing algorithms. Finally, we conduct experiments based on real-world dataset, to compare the Restless-UCB policy with state-of-the-art benchmarks. Our results show that Restless-UCB outperforms existing algorithms in regret, and significantly reduces the running time.

[Predictive Information Accelerates Learning in RL](#)

- Kuang-Huei Lee, Ian Fischer, Anthony Liu, Yijie Guo, Honglak Lee, John Canny, Sergio Guadarrama
- abstract@[open-review](#): The Predictive Information is the mutual information between the past and the future, $I(X_{\text{past}}; X_{\text{future}})$. We hypothesize that capturing the predictive information is useful in RL, since the ability to model what will happen next is necessary for success on many tasks. To test our hypothesis, we train Soft Actor-Critic (SAC) agents from pixels with an auxiliary task that learns a compressed representation of the predictive information of the RL environment dynamics using a contrastive version of the Conditional Entropy Bottleneck (CEB) objective. We refer to these as Predictive Information SAC (PI-SAC) agents. We show that PI-SAC agents can substantially improve sample efficiency over challenging baselines on tasks from the DM Control suite of continuous control environments. We evaluate PI-SAC agents by comparing against uncompressed PI-SAC agents, other compressed and uncompressed agents, and SAC agents directly trained from pixels. Our implementation is given on GitHub.

[Robust and Heavy-Tailed Mean Estimation Made Simple, via Regret Minimization](#)

- Sam Hopkins, Jerry Li, Fred Zhang
- abstract@[open-review](#): We study the problem of estimating the mean of a distribution in high dimensions when either the samples are adversarially corrupted or the distribution is heavy-tailed. Recent developments in robust statistics have established efficient and (near) optimal procedures for both settings. However, the algorithms developed on each side tend to be sophisticated and do not directly transfer to the other, with many of them having ad-hoc or complicated analyses. In this paper, we provide a meta-problem and a duality theorem that lead to a new unified view on robust and heavy-tailed mean estimation in high dimensions. We show that the meta-problem can be solved either by a variant of the Filter algorithm from the recent literature on robust estimation or by the quantum entropy scoring scheme (QUE), due to Dong, Hopkins and Li (NeurIPS '19). By leveraging our duality theorem, these results translate into simple and efficient algorithms for both robust and heavy-tailed settings. Furthermore, the QUE-based procedure has run-time that

matches the fastest known algorithms on both fronts. Our analysis of Filter is through the classic regret bound of the multiplicative weights update method. This connection allows us to avoid the technical complications in previous works and improve upon the run-time analysis of a gradient-descent-based algorithm for robust mean estimation by Cheng, Diakonikolas, Ge and Soltanolkotabi (ICML '20).

[High-Fidelity Generative Image Compression](#)

- Fabian Mentzer, George D. Toderici, Michael Tschannen, Eirikur Agustsson
- abstract@[open-review](#): We extensively study how to combine Generative Adversarial Networks and learned compression to obtain a state-of-the-art generative lossy compression system. In particular, we investigate normalization layers, generator and discriminator architectures, training strategies, as well as perceptual losses. In contrast to previous work, i) we obtain visually pleasing reconstructions that are perceptually similar to the input, ii) we operate in a broad range of bitrates, and iii) our approach can be applied to high-resolution images. We bridge the gap between rate-distortion-perception theory and practice by evaluating our approach both quantitatively with various perceptual metrics, and with a user study. The study shows that our method is preferred to previous approaches even if they use more than 2x the bitrate.

[A Statistical Mechanics Framework for Task-Agnostic Sample Design in Machine Learning](#)

- Bhavya Kailkhura, Jayaraman Thiagarajan, Qunwei Li, Jize Zhang, Yi Zhou, Timo Bremer
- abstract@[open-review](#): In this paper, we present a statistical mechanics framework to understand the effect of sampling properties of training data on the generalization gap of machine learning (ML) algorithms. We connect the generalization gap to the spatial properties of a sample design characterized by the pair correlation function (PCF). In particular, we express generalization gap in terms of the power spectra of the sample design and that of the function to be learned. Using this framework, we show that space-filling sample designs, such as blue noise and Poisson disk sampling, which optimize spectral properties, outperform random designs in terms of the generalization gap and characterize this gain in a closed-form. Our analysis also sheds light on design principles for constructing optimal task-agnostic sample designs that minimize the generalization gap. We corroborate our findings using regression experiments with neural networks on: a) synthetic functions, and b) a complex scientific simulator for inertial confinement fusion (ICF).

[Counterexample-Guided Learning of Monotonic Neural Networks](#)

- Aishwarya Sivaraman, Golnoosh Farnadi, Todd Millstein, Guy Van den Broeck
- abstract@[open-review](#): The widespread adoption of deep learning is often attributed to its automatic feature construction with minimal inductive bias. However, in many real-world tasks, the learned function is intended to satisfy domain-specific constraints. We focus on monotonicity constraints, which are common and require that the function's output increases with increasing values of specific input features. We develop a counterexample-guided technique to provably enforce monotonicity constraints at prediction time. Additionally, we propose a technique to use monotonicity as an inductive bias for deep learning. It works by iteratively incorporating monotonicity counterexamples in the learning process. Contrary to prior work in monotonic learning, we target general ReLU neural networks and do not further restrict the hypothesis space. We have implemented these techniques in a tool called COMET. Experiments on real-world datasets demonstrate that our approach achieves state-of-the-art results compared to existing monotonic learners, and can improve the model quality compared to those that were trained without taking monotonicity constraints into account.

[A Novel Approach for Constrained Optimization in Graphical Models](#)

- Sara Rouhani, Tahrima Rahman, Vibhav Gogate
- abstract@[open-review](#): We consider the following constrained maximization problem in discrete probabilistic graphical models (PGMs). Given two (possibly identical) PGMs $\$M_1\$$ and $\$M_2\$$ defined over the same set of variables and a real number $\$q\$$, find an assignment of values to all variables such that the probability of the assignment is maximized w.r.t. $\$M_1\$$ and is smaller than $\$q\$$ w.r.t. $\$M_2\$$. We show that several explanation and robust estimation queries over graphical models are special cases of this problem. We propose a class of approximate algorithms for solving this problem. Our algorithms are based on a graph concept called $\$k\$$ -separator and heuristic algorithms for multiple choice knapsack and subset-sum problems. Our experiments show that our algorithms are superior to the following approach: encode the problem as a mixed integer linear program (MILP) and solve the latter using a state-of-the-art MILP solver such as SCIP.

[Global Convergence of Deep Networks with One Wide Layer Followed by Pyramidal Topology](#)

- Quynh N. Nguyen, Marco Mondelli
- abstract@[open-review](#): Recent works have shown that gradient descent can find a global minimum for over-parameterized neural networks where the widths of all the hidden layers scale polynomially with N (N being the number of training samples). In this paper, we prove that, for deep networks, a single layer of width N following the input layer suffices to ensure a similar guarantee. In particular, all the remaining layers are allowed to have constant widths, and form a pyramidal topology. We show an application of our result to the widely used Xavier's initialization and obtain an over-parameterization requirement for the single wide layer of order N^2 .

[On the Trade-off between Adversarial and Backdoor Robustness](#)

- Cheng-Hsin Weng, Yan-Ting Lee, Shan-Hung (Brandon) Wu
- abstract@[open-review](#): Deep neural networks are shown to be susceptible to both adversarial attacks and backdoor attacks. Although many defenses against an individual type of the above attacks have been proposed, the interactions between the vulnerabilities of a network to both types of attacks have not been carefully investigated yet. In this paper, we conduct experiments to study whether adversarial robustness and backdoor robustness can affect each other and find a trade-off by increasing the robustness of a network to adversarial examples, the network becomes more vulnerable to backdoor attacks. We then investigate the cause and show how such a trade-off can be exploited for either good or bad purposes. Our findings suggest that future research on defense should take both adversarial and backdoor attacks into account when designing algorithms or robustness measures to avoid pitfalls and a false sense of security.

[Implicit Graph Neural Networks](#)

- Fangda Gu, Heng Chang, Wenwu Zhu, Somayeh Sojoudi, Laurent El Ghaoui
- abstract@[open-review](#): Graph Neural Networks (GNNs) are widely used deep learning models that learn meaningful representations from graph-structured data. Due to the finite nature of the underlying recurrent structure, current GNN methods may struggle to capture long-range dependencies in underlying graphs. To overcome this difficulty, we propose a graph learning framework, called Implicit Graph Neural Networks (IGNN), where predictions are based on the solution of a fixed-point equilibrium equation involving implicitly defined "state" vectors. We use the Perron-Frobenius theory to derive sufficient conditions that ensure well-posedness of the framework. Leveraging implicit differentiation, we derive a tractable projected gradient descent method to train the framework. Experiments on a comprehensive range of tasks show that IGNNs consistently capture long-range dependencies and outperform state-of-the-art GNN models.

[Rethinking Importance Weighting for Deep Learning under Distribution Shift](#)

- Tongtong Fang, Nan Lu, Gang Niu, Masashi Sugiyama
- abstract@[open-review](#): Under distribution shift (DS) where the training data distribution differs from the test one, a powerful technique is importance weighting (IW) which handles DS in two separate steps: weight estimation (WE) estimates the test-over-training density ratio and weighted classification (WC) trains the classifier from weighted training data. However, IW cannot work well on complex data, since WE is incompatible with deep learning. In this paper, we rethink IW and theoretically show it suffers from a circular dependency: we need not only WE for WC, but also WC for WE where a trained deep classifier is used as the feature extractor (FE). To cut off the dependency, we try to pretrain FE from unweighted training data, which leads to biased FE. To overcome the bias, we propose an end-to-end solution dynamic IW that iterates between WE and WC and combines them in a seamless manner, and hence our WE can also enjoy deep networks and stochastic optimizers indirectly. Experiments with two representative types of DS on three popular datasets show that our dynamic IW compares favorably with state-of-the-art methods.

[Guiding Deep Molecular Optimization with Genetic Exploration](#)

- Sungsoo Ahn, Junsu Kim, Hankook Lee, Jinwoo Shin
- abstract@[open-review](#): De novo molecular design attempts to search over the chemical space for molecules with the desired property. Recently, deep learning has gained considerable attention as a promising approach to solve the problem. In this paper, we propose genetic expert-guided learning (GEGL), a simple yet novel framework for training a deep neural network (DNN) to generate highly-rewarding molecules. Our main idea is to design a "genetic expert improvement" procedure, which generates high-quality targets for imitation learning of the DNN. Extensive experiments show that GEGL significantly improves over state-of-the-art methods. For example, GEGL manages to solve the penalized octanol-water partition coefficient optimization with a score of 31.40, while the best-known score in the literature is 27.22. Besides, for the GuacaMol benchmark with 20 tasks, our method achieves the highest score for 19 tasks, in comparison with state-of-the-art methods, and newly obtains the perfect score for three tasks. Our training code is available at <https://github.com/sungsoo-ahn/genetic-expert-guided-learning>.

[Temporal Spike Sequence Learning via Backpropagation for Deep Spiking Neural Networks](#)

- Wenrui Zhang, Peng Li
- abstract@[open-review](#): Spiking neural networks (SNNs) are well suited for spatio-temporal learning and implementations on energy-efficient event-driven neuromorphic processors. However, existing SNN error backpropagation (BP) methods lack proper handling of spiking discontinuities and suffer from low performance compared with the BP methods for traditional artificial neural networks. In addition, a large number of time steps are typically required to achieve decent performance, leading to high latency and rendering spike-based computation unscalable to deep architectures. We present a novel Temporal Spike Sequence Learning Backpropagation (TSSL-BP) method for training deep SNNs, which breaks down error backpropagation across two types of inter-neuron and intra-neuron dependencies and leads to improved temporal learning precision. It captures inter-neuron dependencies through presynaptic firing times by considering the all-or-none characteristics of firing activities and captures intra-neuron dependencies by handling the internal evolution of each neuronal state in time. TSSL-BP efficiently trains deep SNNs within a much shortened temporal window of a few steps while improving the accuracy for various image classification datasets including CIFAR10.

[TSPNet: Hierarchical Feature Learning via Temporal Semantic Pyramid for Sign Language Translation](#)

- DONGXU LI, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, Hongdong Li
- abstract@[open-review](#): Sign language translation (SLT) aims to interpret sign video sequences into text-based natural language sentences. Sign videos consist of continuous sequences of sign gestures with no clear boundaries in between. Existing SLT models usually represent sign visual features in a frame-wise manner so as to avoid needing to explicitly segmenting the videos into isolated signs. However, these methods neglect the temporal information of signs and lead to substantial ambiguity in translation. In this paper, we explore the temporal semantic structures of sign videos to learn more discriminative features. To this end, we first present a novel sign video segment representation which takes into account multiple temporal granularities, thus alleviating the need for accurate video segmentation. Taking advantage of the proposed segment representation, we develop a novel hierarchical sign video feature learning method via a temporal semantic pyramid network, called TSPNet. Specifically, TSPNet introduces an inter-scale attention to evaluate and enhance local semantic consistency of sign segments and an intra-scale attention to resolve semantic ambiguity by using non-local video context. Experiments show that our TSPNet outperforms the state-of-the-art with significant improvements on the BLEU score (from 9.58 to 13.41) and ROUGE score (from 31.80 to 34.96) on the largest commonly used SLT dataset. Our implementation is available at <https://github.com/verashira/TSPNet>.

[Neural Topographic Factor Analysis for fMRI Data](#)

- Eli Sennesh, Zulqarnain Khan, Yiyu Wang, J Benjamin Hutchinson, Ajay Satpute, Jennifer Dy, Jan-Willem van de Meent
- abstract@[open-review](#): Neuroimaging studies produce gigabytes of spatio-temporal data for a small number of participants and stimuli. Recent work increasingly suggests that the common practice of averaging across participants and stimuli leaves out systematic and meaningful information. We propose Neural Topographic Factor Analysis (NTFA), a probabilistic factor analysis model that infers embeddings for participants and stimuli. These embeddings allow us to reason about differences between participants and stimuli as signal rather than noise. We evaluate NTFA on data from an in-house pilot experiment, as well as two publicly available datasets. We demonstrate that inferring representations for participants and stimuli improves predictive generalization to unseen data when compared to previous topographic methods. We also demonstrate that the inferred latent factor representations are useful for downstream tasks such as multivoxel pattern analysis and functional connectivity.

[Neural Architecture Generator Optimization](#)

- Robin Ru, Pedro EsperanÃ§a, Fabio Maria Carlucci
- abstract@[open-review](#): Neural Architecture Search (NAS) was first proposed to achieve state-of-the-art performance through the discovery of new architecture patterns, without human intervention. An over-reliance on expert knowledge in the search space design has however led to increased performance (local optima) without significant architectural breakthroughs, thus preventing truly novel solutions from being reached. In this work we 1) are the first to investigate casting NAS as a problem of finding the optimal network generator and 2) we propose a new, hierarchical and graph-based search space capable of representing an extremely large variety of network types, yet only requiring few continuous hyper-parameters. This greatly reduces the dimensionality of the problem, enabling the effective use of Bayesian Optimisation as a search strategy. At the same time, we expand the range of valid architectures, motivating a multi-objective learning approach. We demonstrate the effectiveness of this strategy on six benchmark datasets and show that our search space generates extremely lightweight yet highly competitive models.

[A Bandit Learning Algorithm and Applications to Auction Design](#)

- Kim Thang Nguyen
- abstract@[open-review](#): We consider online bandit learning in which at every time step, an algorithm has to make a decision and then observe only its reward. The goal is to design efficient (polynomial-time) algorithms that achieve a total reward approximately close to that of the best fixed decision in hindsight. In this paper, we introduce a new notion of (λ, μ) -concave functions and present a bandit learning algorithm that achieves a performance guarantee which is characterized as a function of the concavity parameters λ and μ . The algorithm is based on the mirror descent algorithm in which the update directions follow the gradient of the multilinear extensions of the reward functions. The regret bound induced by our algorithm is $\widetilde{O}(\sqrt{T})$ which is nearly optimal.

We apply our algorithm to auction design, specifically to welfare maximization, revenue maximization, and no-envy learning in auctions. In welfare maximization, we show that a version of fictitious play in smooth auctions guarantees a competitive regret bound which is determined by the smooth parameters. In revenue maximization, we consider the simultaneous second-price auctions with reserve prices in multi-parameter environments. We give a bandit algorithm which achieves the total revenue at least \$1/2\$ times that of the best fixed reserve prices in hindsight. In no-envy learning, we study the bandit item selection problem where the player valuation is submodular and provide an efficient \$1/2\$-approximation no-envy algorithm.

[MetaPoison: Practical General-purpose Clean-label Data Poisoning](#)

- W. Ronny Huang, Jonas Geiping, Liam Fowl, Gavin Taylor, Tom Goldstein
- abstract@[open-review](#): Data poisoning---the process by which an attacker takes control of a model by making imperceptible changes to a subset of the training data---is an emerging threat in the context of neural networks. Existing attacks for data poisoning neural networks have relied on hand-crafted heuristics, because solving the poisoning problem directly via bilevel optimization is generally thought of as intractable for deep models. We propose MetaPoison, a first-order method that approximates the bilevel problem via meta-learning and crafts poisons that fool neural networks. MetaPoison is effective: it outperforms previous clean-label poisoning methods by a large margin. MetaPoison is robust: poisoned data made for one model transfer to a variety of victim models with unknown training settings and architectures. MetaPoison is general-purpose, it works not only in fine-tuning scenarios, but also for end-to-end training from scratch, which till now hasn't been feasible for clean-label attacks with deep nets. MetaPoison can achieve arbitrary adversary goals---like using poisons of one class to make a target image don the label of another arbitrarily chosen class. Finally, MetaPoison works in the real-world. We demonstrate for the first time successful data poisoning of models trained on the black-box Google Cloud AutoML API.

[Sample Efficient Reinforcement Learning via Low-Rank Matrix Estimation](#)

- Devavrat Shah, Dogyoon Song, Zhi Xu, Yuzhe Yang
- abstract@[open-review](#): We consider the question of learning $\$Q\$$ -function in a sample efficient manner for reinforcement learning with continuous state and action spaces under a generative model. If $\$Q\$$ -function is Lipschitz continuous, then the minimal sample complexity for estimating $\$epsilon$$ -optimal $\$Q\$$ -function is known to scale as $\$Omega(frac{1}{epsilon^{d_1+d_2+2}})$ per classical non-parametric learning theory, where d_1 and d_2 denote the dimensions of the state and action spaces respectively. The $\$Q\$$ -function, when viewed as a kernel, induces a Hilbert-Schmidt operator and hence possesses square-summable spectrum. This motivates us to consider a parametric class of $\$Q\$$ -functions parameterized by its "rank" r , which contains all Lipschitz $\$Q\$$ -functions as $\$r \rightarrow \infty$. As our key contribution, we develop a simple, iterative learning algorithm that finds $\$epsilon$$ -optimal $\$Q\$$ -function with sample complexity of $\$widetilde{O}(frac{1}{epsilon^{max(d_1, d_2)+2}})$ when the optimal $\$Q\$$ -function has low rank r and the discounting factor γ is below a certain threshold. Thus, this provides an exponential improvement in sample complexity. To enable our result, we develop a novel Matrix Estimation algorithm that faithfully estimates an unknown low-rank matrix in the $\$ell_\infty$ sense even in the presence of arbitrary bounded noise, which might be of interest in its own right. Empirical results on several stochastic control tasks confirm the efficacy of our "low-rank" algorithms.

[Training Generative Adversarial Networks with Limited Data](#)

- Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, Timo Aila
- abstract@[open-review](#): Training generative adversarial networks (GAN) using too little data typically leads to discriminator overfitting, causing training to diverge. We propose an adaptive discriminator augmentation mechanism that significantly stabilizes training in limited data regimes. The approach does not require changes to loss functions or network architectures, and is applicable both when training from scratch and when fine-tuning an existing GAN on another dataset. We demonstrate, on several datasets, that good results are now possible using only a few thousand training images, often matching StyleGAN2 results with an order of magnitude fewer images. We expect this to open up new application domains for GANs. We also find that the widely used CIFAR-10 is, in fact, a limited data benchmark, and improve the record FID from 5.59 to 2.42.

[Deeply Learned Spectral Total Variation Decomposition](#)

- Tamara G. Grossmann, Yury Korolev, Guy Gilboa, Carola Schoenlieb
- abstract@[open-review](#): Non-linear spectral decompositions of images based on one-homogeneous functionals such as total variation have gained considerable attention in the last few years. Due to their ability to extract spectral components corresponding to objects of different size and contrast, such decompositions enable filtering, feature transfer, image fusion and other applications. However, obtaining this decomposition involves solving multiple non-smooth optimisation problems and is therefore computationally highly intensive. In this paper, we present a neural network approximation of a non-linear spectral decomposition. We report up to four orders of magnitude ($\sim 10,000$) speedup in processing of mega-pixel size images, compared to classical GPU implementations. Our proposed network, TVspecNET, is able to implicitly learn the underlying PDE and, despite being entirely data driven, inherits invariances of the model based transform. To the best of our knowledge, this is the first approach towards learning a non-linear spectral decomposition of images. Not only do we gain a staggering computational advantage, but this approach can also be seen as a step towards studying neural networks that can decompose an image into spectral components defined by a user rather than a handcrafted functional.

[FracTrain: Fractionally Squeezing Bit Savings Both Temporally and Spatially for Efficient DNN Training](#)

- Yonggan Fu, Haoran You, Yang Zhao, Yue Wang, Chaojian Li, Kailash Gopalakrishnan, Zhangyang Wang, Yingyan Lin
- abstract@[open-review](#): Recent breakthroughs in deep neural networks (DNNs) have fueled a tremendous demand for intelligent edge devices featuring on-site learning, while the practical realization of such systems remains a challenge due to the limited resources available at the edge and the required massive training costs for state-of-the-art (SOTA) DNNs. As reducing precision is one of the most effective knobs for boosting training time/energy efficiency, there has been a growing interest in low-precision DNN training. In this paper, we explore from an orthogonal direction: how to fractionally squeeze out more training cost savings from the most redundant bit level, progressively along the training trajectory and dynamically per input. Specifically, we propose FracTrain that integrates (i) progressive fractional quantization which gradually increases the precision of activations, weights, and gradients that will not reach the precision of SOTA static quantized DNN training until the final training stage, and (ii) dynamic fractional quantization which assigns precisions to both the activations and gradients of each layer in an input-adaptive manner, for only "fractionally" updating layer parameters. Extensive simulations and ablation studies (six models, four datasets, and three training settings including standard, adaptation, and fine-tuning) validate the effectiveness of FracTrain in reducing computational cost and hardware-quantified energy-latency of DNN training while achieving a comparable or better (-0.12%~+1.87%) accuracy. For example, when training ResNet-74 on CIFAR-10, FracTrain achieves 77.6% and 53.5% computational cost and training latency savings, respectively, compared with the best SOTA baseline, while achieving a comparable (-0.07%) accuracy. Our codes are available at: <https://github.com/RICE-EIC/FracTrain>.

[Improving Neural Network Training in Low Dimensional Random Bases](#)

- Frithjof Gressmann, Zach Eaton-Rosen, Carlo Luschi
- abstract@[open-review](#): Stochastic Gradient Descent (SGD) has proven to be remarkably effective in optimizing deep neural networks that employ ever-larger numbers of parameters. Yet, improving the efficiency of large-scale optimization remains a vital and highly active area of research. Recent work has shown that deep neural networks can be optimized in randomly-projected subspaces of much smaller dimensionality than their native parameter space. While such training is promising for more efficient and scalable optimization schemes, its practical application is limited by inferior optimization performance. Here, we improve on recent random subspace approaches as follows. We show that keeping the random projection fixed throughout training

is detrimental to optimization. We propose re-drawing the random subspace at each step, which yields significantly better performance. We realize further improvements by applying independent projections to different parts of the network, making the approximation more efficient as network dimensionality grows. To implement these experiments, we leverage hardware-accelerated pseudo-random number generation to construct the random projections on-demand at every optimization step, allowing us to distribute the computation of independent random directions across multiple workers with shared random seeds. This yields significant reductions in memory and is up to 10x faster for the workloads in question.

[Safe Reinforcement Learning via Curriculum Induction](#)

- Matteo Turchetta, Andrey Kolobov, Shital Shah, Andreas Krause, Alekh Agarwal
- abstract@[open-review](#): In safety-critical applications, autonomous agents may need to learn in an environment where mistakes can be very costly. In such settings, the agent needs to behave safely not only after but also while learning. To achieve this, existing safe reinforcement learning methods make an agent rely on priors that let it avoid dangerous situations during exploration with high probability, but both the probabilistic guarantees and the smoothness assumptions inherent in the priors are not viable in many scenarios of interest such as autonomous driving. This paper presents an alternative approach inspired by human teaching, where an agent learns under the supervision of an automatic instructor that saves the agent from violating constraints during learning. In this model, we introduce the monitor that neither needs to know how to do well at the task the agent is learning nor needs to know how the environment works. Instead, it has a library of reset controllers that it activates when the agent starts behaving dangerously, preventing it from doing damage. Crucially, the choices of which reset controller to apply in which situation affect the speed of agent learning. Based on observing agents' progress the teacher itself learns a policy for choosing the reset controllers, a curriculum, to optimize the agent's final policy reward. Our experiments use this framework in two environments to induce curricula for safe and efficient learning.

[Leverage the Average: an Analysis of KL Regularization in Reinforcement Learning](#)

- Nino Vieillard, Tadashi Kozuno, Bruno Scherrer, Olivier Pietquin, Remi Munos, Matthieu Geist
- abstract@[open-review](#): Recent Reinforcement Learning (RL) algorithms making use of Kullback-Leibler (KL) regularization as a core component have shown outstanding performance. Yet, only little is understood theoretically about why KL regularization helps, so far. We study KL regularization within an approximate value iteration scheme and show that it implicitly averages q-values. Leveraging this insight, we provide a very strong performance bound, the very first to combine two desirable aspects: a linear dependency to the horizon (instead of quadratic) and an error propagation term involving an averaging effect of the estimation errors (instead of an accumulation effect). We also study the more general case of an additional entropy regularizer. The resulting abstract scheme encompasses many existing RL algorithms. Some of our assumptions do not hold with neural networks, so we complement this theoretical analysis with an extensive empirical study.

[How Robust are the Estimated Effects of Nonpharmaceutical Interventions against COVID-19?](#)

- Mrinank Sharma, Svenja Mindermann, Jan Brauner, Gavin Leech, Anna Stephenson, Tomáš Gavenčík, Jan Kulveit, Yee Whye Teh, Leonid Chindelevitch, Yarin Gal
- abstract@[open-review](#): To what extent are effectiveness estimates of nonpharmaceutical interventions (NPIs) against COVID-19 influenced by the assumptions our models make? To answer this question, we investigate 2 state-of-the-art NPI effectiveness models and propose 6 variants that make different structural assumptions. In particular, we investigate how well NPI effectiveness estimates generalise to unseen countries, and their sensitivity to unobserved factors. Models which account for noise in disease transmission compare favourably. We further evaluate how robust estimates are to different choices of epidemiological parameters and data. Focusing on models that assume transmission noise, we find that previously published results are robust across these choices and across different models. Finally, we mathematically ground the interpretation of NPI effectiveness estimates when certain common assumptions do not hold.

[Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses](#)

- Kaivalya Rawal, Himabindu Lakkaraju
- abstract@[open-review](#): As predictive models are increasingly being deployed in high-stakes decision-making, there has been a lot of interest in developing algorithms which can provide recourses to affected individuals. While developing such tools is important, it is even more critical to analyze and interpret a predictive model, and yet it thoroughly to ensure that the recourses it offers are meaningful and non-discriminatory before it is deployed in the real world. To this end, we propose a novel model agnostic framework called Actionable Recourse Summaries (AReS) to construct global counterfactual explanations which provide an interpretable and accurate summary of recourses for the entire population. We formulate a novel objective which simultaneously optimizes for correctness of the recourses and interpretability of the explanations, while minimizing overall recourse costs across the entire population. More specifically, our objective enables us to learn, with optimality guarantees on recourse correctness, a small number of compact rule sets each of which capture recourses for well defined subpopulations within the data. We also demonstrate theoretically that several of the prior approaches proposed to generate recourses for individuals are special cases of our framework. Experimental evaluation with real world datasets and user studies demonstrate that our framework can provide decision makers with a comprehensive overview of recourses corresponding to any black box model, and consequently help detect undesirable model biases and discrimination.

[Generalization error in high-dimensional perceptrons: Approaching Bayes error with convex optimization](#)

- Benjamin Aubin, Florent Krzakala, Yue Lu, Lenka Zdeborová
- abstract@[open-review](#): We consider a commonly studied supervised classification of a synthetic dataset whose labels are generated by feeding a one-layer non-linear neural network with random iid inputs. We study the generalization performances of standard classifiers in the high-dimensional regime where $\alpha = \frac{n}{d}$ is kept finite in the limit of a high dimension d and number of samples n . Our contribution is three-fold: First, we prove a formula for the generalization error achieved by ℓ_2 regularized classifiers that minimize a convex loss. This formula was first obtained by the heuristic replica method of statistical physics. Secondly, focussing on commonly used loss functions and optimizing the ℓ_2 regularization strength, we observe that while ridge regression performance is poor, logistic and hinge regression are surprisingly able to approach the Bayes-optimal generalization error extremely closely. As $\alpha \rightarrow \infty$ they lead to Bayes-optimal rates, a fact that does not follow from predictions of margin-based generalization error bounds. Third, we design an optimal loss and regularizer that provably leads to Bayes-optimal generalization error.

[Projection Efficient Subgradient Method and Optimal Nonsmooth Frank-Wolfe Method](#)

- Kiran K. Thekumparampil, Prateek Jain, Praneeth Netrapalli, Sewoong Oh
- abstract@[open-review](#): We consider the classical setting of optimizing a nonsmooth Lipschitz continuous convex function over a convex constraint set, when having access to a (stochastic) first-order oracle (FO) for the function and a projection oracle (PO) for the constraint set. It is well known that to achieve ϵ -suboptimality in high-dimensions, $\Theta(\epsilon^{-2})$ FO calls are necessary. This is achieved by the projected subgradient method (PGD). However, PGD also entails $O(\epsilon^{-2})$ PO calls, which may be computationally costlier than FO calls (e.g. nuclear norm constraints). Improving this PO calls complexity of PGD is largely unexplored, despite the fundamental nature of this problem and extensive literature. We present first such improvement. This only requires a mild assumption that the objective function, when extended to a slightly larger neighborhood of the constraint set, still remains Lipschitz and accessible via FO. In particular, we introduce MOPES method, which carefully combines Moreau-Yosida smoothing and accelerated first-order schemes. This is guaranteed to find a feasible ϵ -suboptimal solution using only $O(\epsilon^{-1})$ PO calls and optimal $O(\epsilon^{-2})$ FO calls. Further, instead of a PO if we only have a linear minimization oracle (LMO, a la Frank-Wolfe) to access

the constraint set, an extension of our method, MOLES, finds a feasible $\$ \backslash epsilon $$ -suboptimal solution using $\$ O(\backslash epsilon^{-2}) \$$ LMO calls and FO calls--both match known lower bounds, resolving a question left open since White (1993). Our experiments confirm that these methods achieve significant speedups over the state-of-the-art, for a problem with costly PO and LMO calls.

[PGM-Explainer: Probabilistic Graphical Model Explanations for Graph Neural Networks](#)

- Minh Vu, My T. Thai
- abstract@[open-review](#): In Graph Neural Networks (GNNs), the graph structure is incorporated into the learning of node representations. This complex structure makes explaining GNNs' predictions become much more challenging. In this paper, we propose PGM-Explainer, a Probabilistic Graphical Model (PGM) model-agnostic explainer for GNNs. Given a prediction to be explained, PGM-Explainer identifies crucial graph components and generates an explanation in form of a PGM approximating that prediction. Different from existing explainers for GNNs where the explanations are drawn from a set of linear functions of explained features, PGM-Explainer is able to demonstrate the dependencies of explained features in form of conditional probabilities. Our theoretical analysis shows that the PGM generated by PGM-Explainer includes the Markov-blanket of the target prediction, i.e. including all its statistical information. We also show that the explanation returned by PGM-Explainer contains the same set of independence statements in the perfect map. Our experiments on both synthetic and real-world datasets show that PGM-Explainer achieves better performance than existing explainers in many benchmark tasks.

[Few-Cost Salient Object Detection with Adversarial-Paced Learning](#)

- Dingwen Zhang, HaiBin Tian, Jungong Han
- abstract@[open-review](#): Detecting and segmenting salient objects from given image scenes has received great attention in recent years. A fundamental challenge in training the existing deep saliency detection models is the requirement of large amounts of annotated data. While gathering large quantities of training data becomes cheap and easy, annotating the data is an expensive process in terms of time, labor and human expertise. To address this problem, this paper proposes to learn the effective salient object detection model based on the manual annotation on a few training images only, thus dramatically alleviating human labor in training models. To this end, we name this new task as the few-cost salient object detection and propose an adversarial-paced learning (APL)-based framework to facilitate the few-cost learning scenario. Essentially, APL is derived from the self-paced learning (SPL) regime but it infers the robust learning pace through the data-driven adversarial learning mechanism rather than the heuristic design of the learning regularizer. Comprehensive experiments on four widely-used benchmark datasets have demonstrated that the proposed approach can effectively approach to the existing supervised deep salient object detection models with only 1k human-annotated training images.

[Minimax Estimation of Conditional Moment Models](#)

- Nishanth Dikkala, Greg Lewis, Lester Mackey, Vasilis Syrgkanis
- abstract@[open-review](#): We develop an approach for estimating models described via conditional moment restrictions, with a prototypical application being non-parametric instrumental variable regression. We introduce a min-max criterion function, under which the estimation problem can be thought of as solving a zero-sum game between a modeler who is optimizing over the hypothesis space of the target model and an adversary who identifies violating moments over a test function space. We analyze the statistical estimation rate of the resulting estimator for arbitrary hypothesis spaces, with respect to an appropriate analogue of the mean squared error metric, for ill-posed inverse problems. We show that when the minimax criterion is regularized with a second moment penalty on the test function and the test function space is sufficiently rich, then the estimation rate scales with the critical radius of the hypothesis and test function spaces, a quantity which typically gives tight fast rates. Our main result follows from a novel localized Rademacher analysis of statistical learning problems defined via minimax objectives. We provide applications of our main results for several hypothesis spaces used in practice such as: reproducing kernel Hilbert spaces, high dimensional sparse linear functions, spaces defined via shape constraints, ensemble estimators such as random forests, and neural networks. For each of these applications we provide computationally efficient optimization methods for solving the corresponding minimax problem (e.g. stochastic first-order heuristics for neural networks). In several applications, we show how our modified mean squared error rate, combined with conditions that bound the ill-posedness of the inverse problem, lead to mean squared error rates. We conclude with an extensive experimental analysis of the proposed methods.

[Causal Imitation Learning With Unobserved Confounders](#)

- Junzhe Zhang, Daniel Kumor, Elias Bareinboim
- abstract@[open-review](#): One of the common ways children learn is by mimicking adults. Imitation learning focuses on learning policies with suitable performance from demonstrations generated by an expert, with an unspecified performance measure, and unobserved reward signal. Popular methods for imitation learning start by either directly mimicking the behavior policy of an expert (behavior cloning) or by learning a reward function that prioritizes observed expert trajectories (inverse reinforcement learning). However, these methods rely on the assumption that covariates used by the expert to determine her/his actions are fully observed. In this paper, we relax this assumption and study imitation learning when sensory inputs of the learner and the expert differ. First, we provide a non-parametric, graphical criterion that is complete (both necessary and sufficient) for determining the feasibility of imitation from the combinations of demonstration data and qualitative assumptions about the underlying environment, represented in the form of a causal model. We then show that when such a criterion does not hold, imitation could still be feasible by exploiting quantitative knowledge of the expert trajectories. Finally, we develop an efficient procedure for learning the imitating policy from experts' trajectories.

[Your GAN is Secretly an Energy-based Model and You Should Use Discriminator Driven Latent Sampling](#)

- Tong Che, Ruixiang ZHANG, Jascha Sohl-Dickstein, Hugo Larochelle, Liam Paull, Yuan Cao, Yoshua Bengio
- abstract@[open-review](#): We show that the sum of the implicit generator log-density $\$ \log p_g \$$ of a GAN with the logit score of the discriminator defines an energy function which yields the true data density when the generator is imperfect but the discriminator is optimal, thus making it possible to improve on the typical generator (with implicit density $\$ p_g \$$). To make that practical, we show that sampling from this modified density can be achieved by sampling in latent space according to an energy-based model induced by the sum of the latent prior log-density and the discriminator output score. This can be achieved by running a Langevin MCMC in latent space and then applying the generator function, which we call Discriminator Driven Latent Sampling~(DDLS). We show that DDLS is highly efficient compared to previous methods which work in the high-dimensional pixel space and can be applied to improve on previously trained GANs of many types. We evaluate DDLS on both synthetic and real-world datasets qualitatively and quantitatively. On CIFAR-10, DDLS substantially improves the Inception Score of an off-the-shelf pre-trained SN-GAN~\cite{ngan} from $\$ 8.22 \$$ to $\$ 9.09 \$$ which is even comparable to the class-conditional BigGAN~\cite{biggan} model. This achieves a new state-of-the-art in unconditional image synthesis setting without introducing extra parameters or additional training.

[Learning Black-Box Attackers with Transferable Priors and Query Feedback](#)

- Jiancheng YANG, Yangzhou Jiang, Xiaoyang Huang, Bingbing Ni, Chenglong Zhao
- abstract@[open-review](#): This paper addresses the challenging black-box adversarial attack problem, where only classification confidence of a victim model is available. Inspired by consistency of visual saliency between different vision models, a surrogate model is expected to improve the attack performance via transferability. By combining transferability-based and query-based black-box attack, we propose a surprisingly simple baseline approach (named SimBA++) using the surrogate model, which significantly outperforms several state-of-the-art methods. Moreover, to efficiently utilize the query feedback, we update the surrogate model in a novel learning scheme, named High-Order Gradient Approximation (HOGA). By constructing a high-order

gradient computation graph, we update the surrogate model to approximate the victim model in both forward and backward pass. The SimBA++ and HOGA result in Learnable Black-Box Attack (LeBA), which surpasses previous state of the art by considerable margins: the proposed LeBA significantly reduces queries, while keeping higher attack success rates close to 100% in extensive ImageNet experiments, including attacking vision benchmarks and defensive models. Code is open source at <https://github.com/TrustworthyDL/LeBA>.

[Locally Differentially Private \(Contextual\) Bandits Learning](#)

- Kai Zheng, Tianle Cai, Weiran Huang, Zhenguo Li, Liwei Wang
- abstract@[open-review](#): We study locally differentially private (LDP) bandits learning in this paper. First, we propose simple black-box reduction frameworks that can solve a large family of context-free bandits learning problems with LDP guarantee. Based on our frameworks, we can improve previous best results for private bandits learning with one-point feedback, such as private Bandits Convex Optimization etc, and obtain the first results for Bandits Convex Optimization (BCO) with multi-point feedback under LDP. LDP guarantee and black-box nature make our frameworks more attractive in real applications compared with previous specifically designed and relatively weaker differentially private (DP) algorithms. Further, we also extend our algorithm to Generalized Linear Bandits with regret bound $\tilde{O}(T^{3/4}/\epsilon)$ under (ϵ, δ) -LDP and it is conjectured to be optimal. Note given existing $\Omega(T)$ lower bound for DP contextual linear bandits (Shariff & Sheffet, NeurIPS 2018), our result shows a fundamental difference between LDP and DP for contextual bandits.

[Invertible Gaussian Reparameterization: Revisiting the Gumbel-Softmax](#)

- Andres Potapczynski, Gabriel Loaiza-Ganem, John P. Cunningham
- abstract@[open-review](#): The Gumbel-Softmax is a continuous distribution over the simplex that is often used as a relaxation of discrete distributions. Because it can be readily interpreted and easily reparameterized, it enjoys widespread use. We propose a modular and more flexible family of reparameterizable distributions where Gaussian noise is transformed into a one-hot approximation through an invertible function. This invertible function is composed of a modified softmax and can incorporate diverse transformations that serve different specific purposes. For example, the stick-breaking procedure allows us to extend the reparameterization trick to distributions with countably infinite support, thus enabling the use of our distribution along nonparametric models, or normalizing flows let us increase the flexibility of the distribution. Our construction enjoys theoretical advantages over the Gumbel-Softmax, such as closed form KL, and significantly outperforms it in a variety of experiments. Our code is available at <https://github.com/cunningham-lab/igr>.

[Kernel Based Progressive Distillation for Adder Neural Networks](#)

- Yixing Xu, Chang Xu, Xinghao Chen, Wei Zhang, Chunjing XU, Yunhe Wang
- abstract@[open-review](#): Adder Neural Networks (ANNs) which only contain additions bring us a new way of developing deep neural networks with low energy consumption. Unfortunately, there is an accuracy drop when replacing all convolution filters by adder filters. The main reason here is the optimization difficulty of ANNs using ℓ_1 -norm, in which the estimation of gradient in back propagation is inaccurate. In this paper, we present a novel method for further improving the performance of ANNs without increasing the trainable parameters via a progressive kernel based knowledge distillation (PKKD) method. A convolutional neural network (CNN) with the same architecture is simultaneously initialized and trained as a teacher network, features and weights of ANN and CNN will be transformed to a new space to eliminate the accuracy drop. The similarity is conducted in a higher-dimensional space to disentangle the difference of their distributions using a kernel based method. Finally, the desired ANN is learned based on the information from both the ground-truth and teacher, progressively. The effectiveness of the proposed method for learning ANN with higher performance is then well-verified on several benchmarks. For instance, the ANN-50 trained using the proposed PKKD method obtains a 76.8% top-1 accuracy on ImageNet dataset, which is 0.6% higher than that of the ResNet-50.

[Adversarial Soft Advantage Fitting: Imitation Learning without Policy Optimization](#)

- Paul Barde, Julien Roy, Wonseok Jeon, Joelle Pineau, Chris Pal, Derek Nowrouzezahrai
- abstract@[open-review](#): Adversarial Imitation Learning alternates between learning a discriminator -- which tells apart expert's demonstrations from generated ones -- and a generator's policy to produce trajectories that can fool this discriminator. This alternated optimization is known to be delicate in practice since it compounds unstable adversarial training with brittle and sample-inefficient reinforcement learning. We propose to remove the burden of the policy optimization steps by leveraging a novel discriminator formulation. Specifically, our discriminator is explicitly conditioned on two policies: the one from the previous generator's iteration and a learnable policy. When optimized, this discriminator directly learns the optimal generator's policy. Consequently, our discriminator's update solves the generator's optimization problem for free: learning a policy that imitates the expert does not require an additional optimization loop. This formulation effectively cuts by half the implementation and computational burden of Adversarial Imitation Learning algorithms by removing the Reinforcement Learning phase altogether. We show on a variety of tasks that our simpler approach is competitive to prevalent Imitation Learning methods.

[Agree to Disagree: Adaptive Ensemble Knowledge Distillation in Gradient Space](#)

- Shangchen Du, Shan You, Xiaojie Li, Jianlong Wu, Fei Wang, Chen Qian, Changshui Zhang
- abstract@[open-review](#): Distilling knowledge from an ensemble of teacher models is expected to have a more promising performance than that from a single one. Current methods mainly adopt a vanilla average rule, i.e., to simply take the average of all teacher losses for training the student network. However, this approach treats teachers equally and ignores the diversity among them. When conflicts or competitions exist among teachers, which is common, the inner compromise might hurt the distillation performance. In this paper, we examine the diversity of teacher models in the gradient space and regard the ensemble knowledge distillation as a multi-objective optimization problem so that we can determine a better optimization direction for the training of student network. Besides, we also introduce a tolerance parameter to accommodate disagreement among teachers. In this way, our method can be seen as a dynamic weighting method for each teacher in the ensemble. Extensive experiments validate the effectiveness of our method for both logits-based and feature-based cases.

[The Wasserstein Proximal Gradient Algorithm](#)

- Adil Salim, Anna Korba, Giulia Luise
- abstract@[open-review](#): Wasserstein gradient flows are continuous time dynamics that define curves of steepest descent to minimize an objective function over the space of probability measures (i.e., the Wasserstein space). This objective is typically a divergence w.r.t. a fixed target distribution. In recent years, these continuous time dynamics have been used to study the convergence of machine learning algorithms aiming at approximating a probability distribution. However, the discrete-time behavior of these algorithms might differ from the continuous time dynamics. Besides, although discretized gradient flows have been proposed in the literature, little is known about their minimization power. In this work, we propose a Forward Backward (FB) discretization scheme that can tackle the case where the objective function is the sum of a smooth and a nonsmooth geodesically convex terms. Using techniques from convex optimization and optimal transport, we analyze the FB scheme as a minimization algorithm on the Wasserstein space. More precisely, we show under mild assumptions that the FB scheme has convergence guarantees similar to the proximal gradient algorithm in Euclidean spaces (resp. similar to the associated Wasserstein gradient flow).

Universally Quantized Neural Compression

- Eirikur Agustsson, Lucas Theis
- abstract@[open-review](#): A popular approach to learning encoders for lossy compression is to use additive uniform noise during training as a differentiable approximation to test-time quantization. We demonstrate that a uniform noise channel can also be implemented at test time using universal quantization (Ziv, 1985). This allows us to eliminate the mismatch between training and test phases while maintaining a completely differentiable loss function. Implementing the uniform noise channel is a special case of the more general problem of communicating a sample, which we prove is computationally hard if we do not make assumptions about its distribution. However, the uniform special case is efficient as well as easy to implement and thus of great interest from a practical point of view. Finally, we show that quantization can be obtained as a limiting case of a soft quantizer applied to the uniform noise channel, bridging compression with and without quantization.

Temporal Variability in Implicit Online Learning

- NicolÃ² Campolongo, Francesco Orabona
- abstract@[open-review](#): In the setting of online learning, Implicit algorithms turn out to be highly successful from a practical standpoint. However, the tightest regret analyses only show marginal improvements over Online Mirror Descent. In this work, we shed light on this behavior carrying out a careful regret analysis. We prove a novel static regret bound that depends on the temporal variability of the sequence of loss functions, a quantity which is often encountered when considering dynamic competitors. We show, for example, that the regret can be constant if the temporal variability is constant and the learning rate is tuned appropriately, without the need of smooth losses. Moreover, we present an adaptive algorithm that achieves this regret bound without prior knowledge of the temporal variability and prove a matching lower bound. Finally, we validate our theoretical findings on classification and regression datasets.

Investigating Gender Bias in Language Models Using Causal Mediation Analysis

- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, Stuart Shieber
- abstract@[open-review](#): Many interpretation methods for neural models in natural language processing investigate how information is encoded inside hidden representations. However, these methods can only measure whether the information exists, not whether it is actually used by the model. We propose a methodology grounded in the theory of causal mediation analysis for interpreting which parts of a model are causally implicated in its behavior. The approach enables us to analyze the mechanisms that facilitate the flow of information from input to output through various model components, known as mediators. As a case study, we apply this methodology to analyzing gender bias in pre-trained Transformer language models. We study the role of individual neurons and attention heads in mediating gender bias across three datasets designed to gauge a model's sensitivity to gender bias. Our mediation analysis reveals that gender bias effects are concentrated in specific components of the model that may exhibit highly specialized behavior.

Off-Policy Imitation Learning from Observations

- Zhuangdi Zhu, Kaixiang Lin, Bo Dai, Jiayu Zhou
- abstract@[open-review](#): Learning from Observations (LfO) is a practical reinforcement learning scenario from which many applications can benefit through the reuse of incomplete resources. Compared to conventional imitation learning (IL), LfO is more challenging because of the lack of expert action guidance. In both conventional IL and LfO, distribution matching is at the heart of their foundation. Traditional distribution matching approaches are sample-costly which depend on on-policy transitions for policy learning. Towards sample-efficiency, some off-policy solutions have been proposed, which, however, either lack comprehensive theoretical justifications or depend on the guidance of expert actions. In this work, we propose a sample-efficient LfO approach which enables off-policy optimization in a principled manner. To further accelerate the learning procedure, we regulate the policy update with an inverse action model, which assists distribution matching from the perspective of mode-covering. Extensive empirical results on challenging locomotion tasks indicate that our approach is comparable with state-of-the-art in terms of both sample-efficiency and asymptotic performance.

Escaping Saddle-Point Faster under Interpolation-like Conditions

- Abhishek Roy, Krishnakumar Balasubramanian, Saeed Ghadimi, Prasant Mohapatra
- abstract@[open-review](#): In this paper, we show that under over-parametrization several standard stochastic optimization algorithms escape saddle-points and converge to local-minimizers much faster. One of the fundamental aspects of over-parametrized models is that they are capable of interpolating the training data. We show that, under interpolation-like assumptions satisfied by the stochastic gradients in an over-parametrization setting, the first-order oracle complexity of Perturbed Stochastic Gradient Descent (PSGD) algorithm to reach an $\$epsilon$ -local-minimizer, matches the corresponding deterministic rate of $O(1/\epsilon^2)$. We next analyze Stochastic Cubic-Regularized Newton (SCRN) algorithm under interpolation-like conditions, and show that the oracle complexity to reach an $\$epsilon$ -local-minimizer under interpolation-like conditions, is $O(1/\epsilon^{2.5})$. While this obtained complexity is better than the corresponding complexity of either PSGD, or SCRN without interpolation-like assumptions, it does not match the rate of $O(1/\epsilon^{1.5})$ corresponding to deterministic Cubic-Regularized Newton method. It seems further Hessian-based interpolation-like assumptions are necessary to bridge this gap. We also discuss the corresponding improved complexities in the zeroth-order settings.

MatÃ©rn Gaussian Processes on Riemannian Manifolds

- Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, Marc Deisenroth (he/him)
- abstract@[open-review](#): Gaussian processes are an effective model class for learning unknown functions, particularly in settings where accurately representing predictive uncertainty is of key importance. Motivated by applications in the physical sciences, the widely-used MatÃ©rn class of Gaussian processes has recently been generalized to model functions whose domains are Riemannian manifolds, by re-expressing said processes as solutions of stochastic partial differential equations. In this work, we propose techniques for computing the kernels of these processes on compact Riemannian manifolds via spectral theory of the Laplace-Beltrami operator in a fully constructive manner, thereby allowing them to be trained via standard scalable techniques such as inducing point methods. We also extend the generalization from the MatÃ©rn to the widely-used squared exponential Gaussian process. By allowing Riemannian MatÃ©rn Gaussian processes to be trained using well-understood techniques, our work enables their use in mini-batch, online, and non-conjugate settings, and makes them more accessible to machine learning practitioners.

Improved Techniques for Training Score-Based Generative Models

- Yang Song, Stefano Ermon
- abstract@[open-review](#): Score-based generative models can produce high quality image samples comparable to GANs, without requiring adversarial optimization. However, existing training procedures are limited to images of low resolution (typically below 32 x 32), and can be unstable under some settings. We provide a new theoretical analysis of learning and sampling from score models in high dimensional spaces, explaining existing failure modes and motivating new solutions that generalize across datasets. To enhance stability, we also propose to maintain an exponential moving average of model weights. With these improvements, we can effortlessly scale score-based generative models to images with unprecedented resolutions ranging from 64 x 64 to 256 x 256. Our score-based models can generate high-fidelity samples that rival best-in-class GANs on various image datasets, including CelebA, FFHQ, and multiple LSUN categories.

[wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations](#)

- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, Michael Auli
- abstract@[open-review](#): We show for the first time that learning powerful representations from speech audio alone followed by fine-tuning on transcribed speech can outperform the best semi-supervised methods while being conceptually simpler. wav2vec 2.0 masks the speech input in the latent space and solves a contrastive task defined over a quantization of the latent representations which are jointly learned. Experiments using all labeled data of LibriSpeech achieve 1.8/3.3 WER on the clean/other test sets. When lowering the amount of labeled data to one hour, wav2vec 2.0 outperforms the previous state of the art on the 100 hour subset while using 100 times less labeled data. Using just ten minutes of labeled data and pre-training on 53k hours of unlabeled data still achieves 4.8/8.2 WER. This demonstrates the feasibility of speech recognition with limited amounts of labeled data.

[A Maximum-Entropy Approach to Off-Policy Evaluation in Average-Reward MDPs](#)

- Nevena Lazic, Dong Yin, Mehrdad Farajtabar, Nir Levine, Dilan Gorur, Chris Harris, Dale Schuurmans
- abstract@[open-review](#): This work focuses on off-policy evaluation (OPE) with function approximation in infinite-horizon undiscounted Markov decision processes (MDPs). For MDPs that are ergodic and linear (i.e. where rewards and dynamics are linear in some known features), we provide the first finite-sample OPE error bound, extending the existing results beyond the episodic and discounted cases. In a more general setting, when the feature dynamics are approximately linear and for arbitrary rewards, we propose a new approach for estimating stationary distributions with function approximation. We formulate this problem as finding the maximum-entropy distribution subject to matching feature expectations under empirical dynamics. We show that this results in an exponential-family distribution whose sufficient statistics are the features, paralleling maximum-entropy approaches in supervised learning. We demonstrate the effectiveness of the proposed OPE approaches in multiple environments.

[Instead of Rewriting Foreign Code for Machine Learning, Automatically Synthesize Fast Gradients](#)

- William Moses, Valentin Churavy
- abstract@[open-review](#): Applying differentiable programming techniques and machine learning algorithms to foreign programs requires developers to either rewrite their code in a machine learning framework, or otherwise provide derivatives of the foreign code. This paper presents Enzyme, a high-performance automatic differentiation (AD) compiler plugin for the LLVM compiler framework capable of synthesizing gradients of statically analyzable programs expressed in the LLVM intermediate representation (IR). Enzyme synthesizes gradients for programs written in any language whose compiler targets LLVM IR including C, C++, Fortran, Julia, Rust, Swift, MLIR, etc., thereby providing native AD capabilities in these languages. Unlike traditional source-to-source and operator-overloading tools, Enzyme performs AD on optimized IR. On a machine-learning focused benchmark suite including Microsoft's ADBench, AD on optimized IR achieves a geometric mean speedup of 4.2 times over AD on IR before optimization allowing Enzyme to achieve state-of-the-art performance. Packaging Enzyme for PyTorch and TensorFlow provides convenient access to gradients of foreign code with state-of-the-art performance, enabling foreign code to be directly incorporated into existing machine learning workflows.

[Does Unsupervised Architecture Representation Learning Help Neural Architecture Search?](#)

- Shen Yan, Yu Zheng, Wei Ao, Xiao Zeng, Mi Zhang
- abstract@[open-review](#): Existing Neural Architecture Search (NAS) methods either encode neural architectures using discrete encodings that do not scale well, or adopt supervised learning-based methods to jointly learn architecture representations and optimize architecture search on such representations which incurs search bias. Despite the widespread use, architecture representations learned in NAS are still poorly understood. We observe that the structural properties of neural architectures are hard to preserve in the latent space if architecture representation learning and search are coupled, resulting in less effective search performance. In this work, we find empirically that pre-training architecture representations using only neural architectures without their accuracies as labels improves the downstream architecture search efficiency. To explain this finding, we visualize how unsupervised architecture representation learning better encourages neural architectures with similar connections and operators to cluster together. This helps map neural architectures with similar performance to the same regions in the latent space and makes the transition of architectures in the latent space relatively smooth, which considerably benefits diverse downstream search strategies.

[Value-driven Hindsight Modelling](#)

- Arthur Guez, Fabio Viola, Theophane Weber, Lars Buesing, Steven Kapturowski, Doina Precup, David Silver, Nicolas Heess
- abstract@[open-review](#): Value estimation is a critical component of the reinforcement learning (RL) paradigm. The question of how to effectively learn value predictors from data is one of the major problems studied by the RL community, and different approaches exploit structure in the problem domain in different ways. Model learning can make use of the rich transition structure present in sequences of observations, but this approach is usually not sensitive to the reward function. In contrast, model-free methods directly leverage the quantity of interest from the future, but receive a potentially weak scalar signal (an estimate of the return). We develop an approach for representation learning in RL that sits in between these two extremes: we propose to learn what to model in a way that can directly help value prediction. To this end, we determine which features of the future trajectory provide useful information to predict the associated return. This provides tractable prediction targets that are directly relevant for a task, and can thus accelerate learning the value function. The idea can be understood as reasoning, in hindsight, about which aspects of the future observations could help past value prediction. We show how this can help dramatically even in simple policy evaluation settings. We then test our approach at scale in challenging domains, including on 57 Atari 2600 games.

[Dynamic Regret of Convex and Smooth Functions](#)

- Peng Zhao, Yu-Jie Zhang, Lijun Zhang, Zhi-Hua Zhou
- abstract@[open-review](#): We investigate online convex optimization in non-stationary environments and choose the dynamic regret as the performance measure, defined as the difference between cumulative loss incurred by the online algorithm and that of any feasible comparator sequence. Let T be the time horizon and P_T be the path-length that essentially reflects the non-stationarity of environments, the state-of-the-art dynamic regret is $\mathcal{O}(\sqrt{T(1+P_T)})$. Although this bound is proved to be minimax optimal for convex functions, in this paper, we demonstrate that it is possible to further enhance the dynamic regret by exploiting the smoothness condition. Specifically, we propose novel online algorithms that are capable of leveraging smoothness and replace the dependence on T in the dynamic regret by problem-dependent quantities: the variation in gradients of loss functions, the cumulative loss of the comparator sequence, and the minimum of the previous two terms. These quantities are at most $\mathcal{O}(T)$ while could be much smaller in benign environments. Therefore, our results are adaptive to the intrinsic difficulty of the problem, since the bounds are tighter than existing results for easy problems and meanwhile guarantee the same rate in the worst case.

[On Convergence of Nearest Neighbor Classifiers over Feature Transformations](#)

- Luka Rimanic, Cedric Renggli, Bo Li, Ce Zhang
- abstract@[open-review](#): The k-Nearest Neighbors (kNN) classifier is a fundamental non-parametric machine learning algorithm. However, it is well known that it suffers from the curse of dimensionality, which is why in practice one often applies a kNN classifier on top of a (pre-trained) feature transformation. From a theoretical perspective, most, if not all theoretical results aimed at understanding the kNN classifier are derived for the raw feature space. This leads to an emerging gap between our theoretical understanding of kNN and its practical applications. In this paper, we take a first step towards bridging

this gap. We provide a novel analysis on the convergence rates of a kNN classifier over transformed features. This analysis requires in-depth understanding of the properties that connect both the transformed space and the raw feature space. More precisely, we build our convergence bound upon two key properties of the transformed space: (1) safety -- how well can one recover the raw posterior from the transformed space, and (2) smoothness -- how complex this recovery function is. Based on our result, we are able to explain why some (pre-trained) feature transformations are better suited for a kNN classifier than other. We empirically validate that both properties have an impact on the kNN convergence on 30 feature transformations with 6 benchmark datasets spanning from the vision to the text domain.

[Mitigating Manipulation in Peer Review via Randomized Reviewer Assignments](#)

- Steven Jecmen, Hanrui Zhang, Ryan Liu, Nihar Shah, Vincent Conitzer, Fei Fang
- abstract@[open-review](#): We consider three important challenges in conference peer review: (i) reviewers maliciously attempting to get assigned to certain papers to provide positive reviews, possibly as part of quid-pro-quo arrangements with the authors; (ii) "torpedo reviewing," where reviewers deliberately attempt to get assigned to certain papers that they dislike in order to reject them; (iii) reviewer de-anonymization on release of the similarities and the reviewer-assignment code. On the conceptual front, we identify connections between these three problems and present a framework that brings all these challenges under a common umbrella. We then present a (randomized) algorithm for reviewer assignment that can optimally solve the reviewer-assignment problem under any given constraints on the probability of assignment for any reviewer-paper pair. We further consider the problem of restricting the joint probability that certain suspect pairs of reviewers are assigned to certain papers, and show that this problem is NP-hard for arbitrary constraints on these joint probabilities but efficiently solvable for a practical special case. Finally, we experimentally evaluate our algorithms on datasets from past conferences, where we observe that they can limit the chance that any malicious reviewer gets assigned to their desired paper to 50% while producing assignments with over 90% of the total optimal similarity.

[Contrastive learning of global and local features for medical image segmentation with limited annotations](#)

- Krishna Chaitanya, Ertunc Erdil, Neerav Karani, Ender Konukoglu
- abstract@[open-review](#): A key requirement for the success of supervised deep learning is a large labeled dataset - a condition that is difficult to meet in medical image analysis. Self-supervised learning (SSL) can help in this regard by providing a strategy to pre-train a neural network with unlabeled data, followed by fine-tuning for a downstream task with limited annotations. Contrastive learning, a particular variant of SSL, is a powerful technique for learning image-level representations. In this work, we propose strategies for extending the contrastive learning framework for segmentation of volumetric medical images in the semi-supervised setting with limited annotations, by leveraging domain-specific and problem-specific cues. Specifically, we propose (1) novel contrasting strategies that leverage structural similarity across volumetric medical images (domain-specific cue) and (2) a local version of the contrastive loss to learn distinctive representations of local regions that are useful for per-pixel segmentation (problem-specific cue). We carry out an extensive evaluation on three Magnetic Resonance Imaging (MRI) datasets. In the limited annotation setting, the proposed method yields substantial improvements compared to other self-supervision and semi-supervised learning techniques. When combined with a simple data augmentation technique, the proposed method reaches within 8% of benchmark performance using only two labeled MRI volumes for training. The code is made public at https://github.com/krishnabits001/domain_specific_cl.

[Self-Supervised Graph Transformer on Large-Scale Molecular Data](#)

- Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying WEI, Wenbing Huang, Junzhou Huang
- abstract@[open-review](#): How to obtain informative representations of molecules is a crucial prerequisite in AI-driven drug design and discovery. Recent researches abstract molecules as graphs and employ Graph Neural Networks (GNNs) for molecular representation learning. Nevertheless, two issues impede the usage of GNNs in real scenarios: (1) insufficient labeled molecules for supervised training; (2) poor generalization capability to new-synthesized molecules. To address them both, we propose a novel framework, GROVER, which stands for Graph Representation frOm self-supervised mEssage passing tRansformer. With carefully designed self-supervised tasks in node-, edge- and graph-level, GROVER can learn rich structural and semantic information of molecules from enormous unlabelled molecular data. Rather, to encode such complex information, GROVER integrates Message Passing Networks into the Transformer-style architecture to deliver a class of more expressive encoders of molecules. The flexibility of GROVER allows it to be trained efficiently on large-scale molecular dataset without requiring any supervision, thus being immunized to the two issues mentioned above. We pre-train GROVER with 100 million parameters on 10 million unlabelled molecules--the biggest GNN and the largest training dataset in molecular representation learning. We then leverage the pre-trained GROVER for molecular property prediction followed by task-specific fine-tuning, where we observe a huge improvement (more than 6% on average) from current state-of-the-art methods on 11 challenging benchmarks. The insights we gained are that well-designed self-supervision losses and largely-expansive pre-trained models enjoy the significant potential on performance boosting.

[Generative Neurosymbolic Machines](#)

- Jindong Jiang, Sungjin Ahn
- abstract@[open-review](#): Reconciling symbolic and distributed representations is a crucial challenge that can potentially resolve the limitations of current deep learning. Remarkable advances in this direction have been achieved recently via generative object-centric representation models. While learning a recognition model that infers object-centric symbolic representations like bounding boxes from raw images in an unsupervised way, no such model can provide another important ability of a generative model, i.e., generating (sampling) according to the structure of learned world density. In this paper, we propose Generative Neurosymbolic Machines, a generative model that combines the benefits of distributed and symbolic representations to support both structured representations of symbolic components and density-based generation. These two crucial properties are achieved by a two-layer latent hierarchy with the global distributed latent for flexible density modeling and the structured symbolic latent map. To increase the model flexibility in this hierarchical structure, we also propose the StructDRAW prior. In experiments, we show that the proposed model significantly outperforms the previous structured representation models as well as the state-of-the-art non-structured generative models in terms of both structure accuracy and image generation quality.

[How many samples is a good initial point worth in Low-rank Matrix Recovery?](#)

- Jialun Zhang, Richard Zhang
- abstract@[open-review](#): Given a sufficiently large amount of labeled data, the nonconvex low-rank matrix recovery problem contains no spurious local minima, so a local optimization algorithm is guaranteed to converge to a global minimum starting from any initial guess. However, the actual amount of data needed by this theoretical guarantee is very pessimistic, as it must prevent spurious local minima from existing anywhere, including at adversarial locations. In contrast, prior work based on good initial guesses have more realistic data requirements, because they allow spurious local minima to exist outside of a neighborhood of the solution. In this paper, we quantify the relationship between the quality of the initial guess and the corresponding reduction in data requirements. Using the restricted isometry constant as a surrogate for sample complexity, we compute a sharp threshold number of samples needed to prevent each specific point on the optimization landscape from becoming a spurious local minima. Optimizing the threshold over regions of the landscape, we see that, for initial points not too close to the ground truth, a linear improvement in the quality of the initial guess amounts to a constant factor improvement in the sample complexity.

[CSER: Communication-efficient SGD with Error Reset](#)

- Cong Xie, Shuai Zheng, Sanmi Koyejo, Indranil Gupta, Mu Li, Haibin Lin

- abstract@[open-review](#): The scalability of Distributed Stochastic Gradient Descent (SGD) is today limited by communication bottlenecks. We propose a novel SGD variant: \underline{C}ommunication-efficient \underline{S}GD with \underline{E}rror \underline{R}eset, or \underline{CSER}. The key idea in CSER is first a new technique called ``error reset'' that adapts arbitrary compressors for SGD, producing bifurcated local models with periodic reset of resulting local residual errors. Second we introduce partial synchronization for both the gradients and the models, leveraging advantages from them. We prove the convergence of CSER for smooth non-convex problems. Empirical results show that when combined with highly aggressive compressors, the CSER algorithms accelerate the distributed training by nearly \$10\times\$ for CIFAR-100, and by \$4.5\times\$ for ImageNet.

[Efficient estimation of neural tuning during naturalistic behavior](#)

- Edoardo Balzani, Kaushik Lakshminarasimhan, Dora Angelaki, Cristina Savin
- abstract@[open-review](#): Recent technological advances in systems neuroscience have led to a shift away from using simple tasks, with low-dimensional, well-controlled stimuli, towards trying to understand neural activity during naturalistic behavior. However, with the increase in number and complexity of task-relevant features, standard analyses such as estimating tuning functions become challenging. Here, we use a Poisson generalized additive model (P-GAM) with spline nonlinearities and an exponential link function to map a large number of task variables (input stimuli, behavioral outputs, or activity of other neurons, modeled as discrete events or continuous variables) into spike counts. We develop efficient procedures for parameter learning by optimizing a generalized cross-validation score and infer marginal confidence bounds for the contribution of each feature to neural responses. This allows us to robustly identify a minimal set of task features that each neuron is responsive to, circumventing computationally demanding model comparison. We show that our estimation procedure outperforms traditional regularized GLMs in terms of both fit quality and computing time. When applied to neural recordings from monkeys performing a virtual reality spatial navigation task, P-GAM reveals mixed selectivity and preferential coupling between neurons with similar tuning.

[High-recall causal discovery for autocorrelated time series with latent confounders](#)

- Andreas Gerhardus, Jakob Runge
- abstract@[open-review](#): We present a new method for linear and nonlinear, lagged and contemporaneous constraint-based causal discovery from observational time series in the presence of latent confounders. We show that existing causal discovery methods such as FCI and variants suffer from low recall in the autocorrelated time series case and identify low effect size of conditional independence tests as the main reason. Information-theoretical arguments show that effect size can often be increased if causal parents are included in the conditioning sets. To identify parents early on, we suggest an iterative procedure that utilizes novel orientation rules to determine ancestral relationships already during the edge removal phase. We prove that the method is order-independent, and sound and complete in the oracle case. Extensive simulation studies for different numbers of variables, time lags, sample sizes, and further cases demonstrate that our method indeed achieves much higher recall than existing methods for the case of autocorrelated continuous variables while keeping false positives at the desired level. This performance gain grows with stronger autocorrelation. At github.com/jakobrunge/tigramite we provide Python code for all methods involved in the simulation studies.

[Forget About the LiDAR: Self-Supervised Depth Estimators with MED Probability Volumes](#)

- Juan Luis GonzalezBello, Munchurl Kim
- abstract@[open-review](#): Self-supervised depth estimators have recently shown results comparable to the supervised methods on the challenging single image depth estimation (SIDE) task, by exploiting the geometrical relations between target and reference views in the training data. However, previous methods usually learn forward or backward image synthesis, but not depth estimation, as they cannot effectively neglect occlusions between the target and the reference images. Previous works rely on rigid photometric assumptions or on the SIDE network to infer depth and occlusions, resulting in limited performance. On the other hand, we propose a method to "Forget About the LiDAR" (FAL), with Mirrored Exponential Disparity (MED) probability volumes for the training of monocular depth estimators from stereo images. Our MED representation allows us to obtain geometrically inspired occlusion maps with our novel Mirrored Occlusion Module (MOM), which does not impose a learning burden on our FAL-net. Contrary to the previous methods that learn SIDE from stereo pairs by regressing disparity in the linear space, our FAL-net regresses disparity by binning it into the exponential space, which allows for better detection of distant and nearby objects. We define a two-step training strategy for our FAL-net: It is first trained for view synthesis and then fine-tuned for depth estimation with our MOM. Our FAL-net is remarkably light-weight and outperforms the previous state-of-the-art methods with 8\$\times\$ fewer parameters and 3\$\times\$ faster inference speeds on the challenging KITTI dataset. We present extensive experimental results on the KITTI, CityScapes, and Make3D datasets to verify our method's effectiveness. To the authors' best knowledge, the presented method performs the best among all the previous self-supervised methods until now.

[Joint Contrastive Learning with Infinite Possibilities](#)

- Qi Cai, Yu Wang, Yingwei Pan, Ting Yao, Tao Mei
- abstract@[open-review](#): This paper explores useful modifications of the recent development in contrastive learning via novel probabilistic modeling. We derive a particular form of contrastive loss named Joint Contrastive Learning (JCL). JCL implicitly involves the simultaneous learning of an infinite number of query-key pairs, which poses tighter constraints when searching for invariant features. We derive an upper bound on this formulation that allows analytical solutions in an end-to-end training manner. While JCL is practically effective in numerous computer vision applications, we also theoretically unveil the certain mechanisms that govern the behavior of JCL. We demonstrate that the proposed formulation harbors an innate agency that strongly favors similarity within each instance-specific class, and therefore remains advantageous when searching for discriminative features among distinct instances. We evaluate these proposals on multiple benchmarks, demonstrating considerable improvements over existing algorithms. Code is publicly available at: <https://github.com/caiqi/Joint-Contrastive-Learning>.

[Robust Gaussian Covariance Estimation in Nearly-Matrix Multiplication Time](#)

- Jerry Li, Guanghao Ye
- abstract@[open-review](#): Robust covariance estimation is the following, well-studied problem in high dimensional statistics: given \$N\$ samples from a \$d\$-dimensional Gaussian \$\mathcal{N}(\boldsymbol{\mu}, \Sigma)\$, but where an \$\varepsilon\$-fraction of the samples have been arbitrarily corrupted, output \$\widehat{\Sigma}\$ minimizing the total variation distance between \$\mathcal{N}(\boldsymbol{\mu}, \Sigma)\$ and \$\mathcal{N}(\boldsymbol{\mu}, \widehat{\Sigma})\$. This corresponds to learning \$\Sigma\$ in a natural affine-invariant variant of the Frobenius norm known as the Mahalanobis norm. Previous work of Cheng et al demonstrated an algorithm that, given \$N = \tilde{\Omega}(d^2/\varepsilon^2)\$ samples, achieved a near-optimal error of \$O(\varepsilon \log(1/\varepsilon))\$, and moreover, their algorithm ran in time \$\tilde{O}(T(N, d) \log \kappa / \text{poly}(\varepsilon))\$, where \$T(N, d)\$ is the time it takes to multiply a \$d \times N\$ matrix by its transpose, and \$\kappa\$ is the condition number of \$\Sigma\$. When \$\varepsilon\$ is relatively small, their polynomial dependence on \$1/\varepsilon\$ in the runtime is prohibitively large. In this paper, we demonstrate a novel algorithm that achieves the same statistical guarantees, but which runs in time \$\tilde{O}(T(N, d) \log \kappa)\$. In particular, our runtime has no dependence on \$\varepsilon\$. When \$\Sigma\$ is reasonably conditioned, our runtime matches that of the fastest algorithm for covariance estimation without outliers, up to poly-logarithmic factors, showing that we can get robustness essentially ``for free.''

[Adversarially-learned Inference via an Ensemble of Discrete Undirected Graphical Models](#)

- Adarsh Keshav Jeewajee, Leslie Kaelbling

- abstract@[open-review](#): Undirected graphical models are compact representations of joint probability distributions over random variables. To solve inference tasks of interest, graphical models of arbitrary topology can be trained using empirical risk minimization. However, to solve inference tasks that were not seen during training, these models (EGMs) often need to be re-trained. Instead, we propose an inference-agnostic adversarial training framework which produces an infinitely-large ensemble of graphical models (AGMs). The ensemble is optimized to generate data within the GAN framework, and inference is performed using a finite subset of these models. AGMs perform comparably with EGMs on inference tasks that the latter were specifically optimized for. Most importantly, AGMs show significantly better generalization to unseen inference tasks compared to EGMs, as well as deep neural architectures like GibbsNet and VAEAC which allow arbitrary conditioning. Finally, AGMs allow fast data sampling, competitive with Gibbs sampling from EGMs.

[GS-WGAN: A Gradient-Sanitized Approach for Learning Differentially Private Generators](#)

- Dingfan Chen, Tribhuvanesh Orekondy, Mario Fritz
- abstract@[open-review](#): The wide-spread availability of rich data has fueled the growth of machine learning applications in numerous domains. However, growth in domains with highly-sensitive data (e.g., medical) is largely hindered as the private nature of data prohibits it from being shared. To this end, we propose Gradient-sanitized Wasserstein Generative Adversarial Networks (GS-WGAN), which allows releasing a sanitized form of the sensitive data with rigorous privacy guarantees. In contrast to prior work, our approach is able to distort gradient information more precisely, and thereby enabling training deeper models which generate more informative samples. Moreover, our formulation naturally allows for training GANs in both centralized and federated (i.e., decentralized) data scenarios. Through extensive experiments, we find our approach consistently outperforms state-of-the-art approaches across multiple metrics (e.g., sample quality) and datasets.

[SurVAE Flows: Surjections to Bridge the Gap between VAEs and Flows](#)

- Didrik Nielsen, Priyank Jaini, Emiel Hoogeboom, Ole Winther, Max Welling
- abstract@[open-review](#): Normalizing flows and variational autoencoders are powerful generative models that can represent complicated density functions. However, they both impose constraints on the models: Normalizing flows use bijective transformations to model densities whereas VAEs learn stochastic transformations that are non-invertible and thus typically do not provide tractable estimates of the marginal likelihood. In this paper, we introduce SurVAE Flows: A modular framework of composable transformations that encompasses VAEs and normalizing flows. SurVAE Flows bridge the gap between normalizing flows and VAEs with surjective transformations, wherein the transformations are deterministic in one direction -- thereby allowing exact likelihood computation, and stochastic in the reverse direction -- hence providing a lower bound on the corresponding likelihood. We show that several recently proposed methods, including dequantization and augmented normalizing flows, can be expressed as SurVAE Flows. Finally, we introduce common operations such as the max value, the absolute value, sorting and stochastic permutation as composable layers in SurVAE Flows.

[Learning Causal Effects via Weighted Empirical Risk Minimization](#)

- Yonghan Jung, Jin Tian, Elias Bareinboim
- abstract@[open-review](#): Learning causal effects from data is a fundamental problem across the sciences. Determining the identifiability of a target effect from a combination of the observational distribution and the causal graph underlying a phenomenon is well-understood in theory. However, in practice, it remains a challenge to apply the identification theory to estimate the identified causal functionals from finite samples. Although a plethora of effective estimators have been developed under the setting known as the back-door (also called conditional ignorability), there exists still no systematic way of estimating arbitrary causal functionals that are both computationally and statistically attractive. This paper aims to bridge this gap, from causal identification to causal estimation. We note that estimating functionals from limited samples based on the empirical risk minimization (ERM) principle has been pervasive in the machine learning literature, and these methods have been extended to causal inference under the back-door setting. In this paper, we develop a learning framework that marries two families of methods, benefiting from the generality of the causal identification theory and the effectiveness of the estimators produced based on the principle of ERM. Specifically, we develop a sound and complete algorithm that generates causal functionals in the form of weighted distributions that are amenable to the ERM optimization. We then provide a practical procedure for learning causal effects from finite samples and a causal graph. Finally, experimental results support the effectiveness of our approach.

[Revisiting the Sample Complexity of Sparse Spectrum Approximation of Gaussian Processes](#)

- Minh Hoang, Nghia Hoang, Hai Pham, David Woodruff
- abstract@[open-review](#): We introduce a new scalable approximation for Gaussian processes with provable guarantees which holds simultaneously over its entire parameter space. Our approximation is obtained from an improved sample complexity analysis for sparse spectrum Gaussian processes (SSGPs). In particular, our analysis shows that under a certain data disentangling condition, an SSGP's prediction and model evidence (for training) can well-approximate those of a full GP with low sample complexity. We also develop a new auto-encoding algorithm that finds a latent space to disentangle latent input coordinates into well-separated clusters, which is amenable to our sample complexity analysis. We validate our proposed method on several benchmarks with promising results supporting our theoretical analysis.

[Incorporating Interpretable Output Constraints in Bayesian Neural Networks](#)

- Wanqian Yang, Lars Lorch, Moritz Graule, Himabindu Lakkaraju, Finale Doshi-Velez
- abstract@[open-review](#): Domains where supervised models are deployed often come with task-specific constraints, such as prior expert knowledge on the ground-truth function, or desiderata like safety and fairness. We introduce a novel probabilistic framework for reasoning with such constraints and formulate a prior that enables us to effectively incorporate them into Bayesian neural networks (BNNs), including a variant that can be amortized over tasks. The resulting Output-Constrained BNN (OC-BNN) is fully consistent with the Bayesian framework for uncertainty quantification and is amenable to black-box inference. Unlike typical BNN inference in uninterpretable parameter space, OC-BNNs widen the range of functional knowledge that can be incorporated, especially for model users without expertise in machine learning. We demonstrate the efficacy of OC-BNNs on real-world datasets, spanning multiple domains such as healthcare, criminal justice, and credit scoring.

[Multi-Stage Influence Function](#)

- Hongge Chen, Si Si, Yang Li, Ciprian Chelba, Sanjiv Kumar, Duane Boning, Cho-Jui Hsieh
- abstract@[open-review](#): Multi-stage training and knowledge transfer, from a large-scale pretraining task to various finetuning tasks, have revolutionized natural language processing and computer vision resulting in state-of-the-art performance improvements. In this paper, we develop a multi-stage influence function score to track predictions from a finetuned model all the way back to the pretraining data. With this score, we can identify the pretraining examples in the pretraining task that contribute most to a prediction in the finetuning task. The proposed multi-stage influence function generalizes the original influence function for a single model in (Koh & Liang, 2017), thereby enabling influence computation through both pretrained and finetuned models. We study two different scenarios with the pretrained embedding fixed or updated in the finetuning tasks. We test our proposed method in various experiments to show its effectiveness and potential applications.

[Probabilistic Fair Clustering](#)

- Seyed Esmaeili, Brian Brubach, Leonidas Tsepenekas, John Dickerson
- abstract@[open-review](#): In clustering problems, a central decision-maker is given a complete metric graph over vertices and must provide a clustering of vertices that minimizes some objective function. In fair clustering problems, vertices are endowed with a color (e.g., membership in a group), and the requirements of a valid clustering might also include the representation of colors in the solution. Prior work in fair clustering assumes complete knowledge of group membership. In this paper, we generalize this by assuming imperfect knowledge of group membership through probabilistic assignments, and present algorithms in this more general setting with approximation ratio guarantees. We also address the problem of "metric membership", where group membership has a notion of order and distance. Experiments are conducted using our proposed algorithms as well as baselines to validate our approach, and also surface nuanced concerns when group membership is not known deterministically.

Stochastic Segmentation Networks: Modelling Spatially Correlated Aleatoric Uncertainty

- Miguel Monteiro, Loic Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, Ben Glocker
- abstract@[open-review](#): In image segmentation, there is often more than one plausible solution for a given input. In medical imaging, for example, experts will often disagree about the exact location of object boundaries. Estimating this inherent uncertainty and predicting multiple plausible hypotheses is of great interest in many applications, yet this ability is lacking in most current deep learning methods. In this paper, we introduce stochastic segmentation networks (SSNs), an efficient probabilistic method for modelling aleatoric uncertainty with any image segmentation network architecture. In contrast to approaches that produce pixel-wise estimates, SSNs model joint distributions over entire label maps and thus can generate multiple spatially coherent hypotheses for a single image. By using a low-rank multivariate normal distribution over the logit space to model the probability of the label map given the image, we obtain a spatially consistent probability distribution that can be efficiently computed by a neural network without any changes to the underlying architecture. We tested our method on the segmentation of real-world medical data, including lung nodules in 2D CT and brain tumours in 3D multimodal MRI scans. SSNs outperform state-of-the-art for modelling correlated uncertainty in ambiguous images while being much simpler, more flexible, and more efficient.

ICE-BeeM: Identifiable Conditional Energy-Based Deep Models Based on Nonlinear ICA

- Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, Aapo Hyvärinen
- abstract@[open-review](#): We consider the identifiability theory of probabilistic models and establish sufficient conditions under which the representations learnt by a very broad family of conditional energy-based models are unique in function space, up to a simple transformation. In our model family, the energy function is the dot-product between two feature extractors, one for the dependent variable, and one for the conditioning variable. We show that under mild conditions, the features are unique up to scaling and permutation. Our results extend recent developments in nonlinear ICA, and in fact, they lead to an important generalization of ICA models. In particular, we show that our model can be used for the estimation of the components in the framework of Independently Modulated Component Analysis (IMCA), a new generalization of nonlinear ICA that relaxes the independence assumption. A thorough empirical study show that representations learnt by our model from real-world image datasets are identifiable, and improve performance in transfer learning and semi-supervised learning tasks.

Testing Determinantal Point Processes

- Khashayar Gatmiry, Maryam Aliakbarpour, Stefanie Jegelka
- abstract@[open-review](#): Determinantal point processes (DPPs) are popular probabilistic models of diversity. In this paper, we investigate DPPs from a new perspective: property testing of distributions. Given sample access to an unknown distribution $\$q\$$ over the subsets of a ground set, we aim to distinguish whether $\$q\$$ is a DPP distribution or $\$epsilon$-far$ from all DPP distributions in $\$ell_1$-distance$. In this work, we propose the first algorithm for testing DPPs. Furthermore, we establish a matching lower bound on the sample complexity of DPP testing. This lower bound also extends to showing a new hardness result for the problem of testing the more general class of log-submodular distributions.

CogLTX: Applying BERT to Long Texts

- Ming Ding, Chang Zhou, Hongxia Yang, Jie Tang
- abstract@[open-review](#): BERTs are incapable of processing long texts due to its quadratically increasing memory and time consumption. The straightforward thoughts to address this problem, such as slicing the text by a sliding window or simplifying transformers, suffer from insufficient long-range attentions or need customized CUDA kernels. The limited text length of BERT reminds us the limited capacity ($5^{1/4} \times 9$ chunks) of the working memory of humans — then how do human beings Cognize Long TeXts? Founded on the cognitive theory stemming from Baddeley, our CogLTX framework identifies key sentences by training a judge model, concatenates them for reasoning and enables multi-step reasoning via rehearsal and decay. Since relevance annotations are usually unavailable, we propose to use treatment experiments to create supervision. As a general algorithm, CogLTX outperforms or gets comparable results to SOTA models on NewsQA, HotpotQA, multi-class and multi-label long-text classification tasks with memory overheads independent of the text length.

f-GAIL: Learning f-Divergence for Generative Adversarial Imitation Learning

- Xin Zhang, Yanhua Li, Ziming Zhang, Zhi-Li Zhang
- abstract@[open-review](#): Imitation learning (IL) aims to learn a policy from expert demonstrations that minimizes the discrepancy between the learner and expert behaviors. Various imitation learning algorithms have been proposed with different pre-determined divergences to quantify the discrepancy. This naturally gives rise to the following question: Given a set of expert demonstrations, which divergence can recover the expert policy more accurately with higher data efficiency? In this work, we propose f-GAIL — a new generative adversarial imitation learning model — that automatically learns a discrepancy measure from the f-divergence family as well as a policy capable of producing expert-like behaviors. Compared with IL baselines with various predefined divergence measures, f-GAIL learns better policies with higher data efficiency in six physics-based control tasks.

Non-parametric Models for Non-negative Functions

- Ulysse Marteau-Ferey, Francis Bach, Alessandro Rudi
- abstract@[open-review](#): Linear models have shown great effectiveness and flexibility in many fields such as machine learning, signal processing and statistics. They can represent rich spaces of functions while preserving the convexity of the optimization problems where they are used, and are simple to evaluate, differentiate and integrate. However, for modeling non-negative functions, which are crucial for unsupervised learning, density estimation, or non-parametric Bayesian methods, linear models are not applicable directly. Moreover, current state-of-the-art models like generalized linear models either lead to non-convex optimization problems, or cannot be easily integrated. In this paper we provide the first model for non-negative functions which benefits from the same good properties of linear models. In particular, we prove that it admits a representer theorem and provide an efficient dual formulation for convex problems. We study its representation power, showing that the resulting space of functions is strictly richer than that of generalized linear models. Finally we extend the model and the theoretical results to functions with outputs in convex cones. The paper is complemented by an experimental evaluation of the model showing its effectiveness in terms of formulation, algorithmic derivation and practical results on the problems of density estimation, regression with heteroscedastic errors, and multiple quantile regression.

Uncertainty Aware Semi-Supervised Learning on Graph Data

- Xujiang Zhao, Feng Chen, Shu Hu, Jin-Hee Cho
- abstract@[open-review](#): Thanks to graph neural networks (GNNs), semi-supervised node classification has shown the state-of-the-art performance in graph data. However, GNNs have not considered different types of uncertainties associated with class probabilities to minimize risk of increasing misclassification under uncertainty in real life. In this work, we propose a multi-source uncertainty framework using a GNN that reflects various types of predictive uncertainties in both deep learning and belief/evidence theory domains for node classification predictions. By collecting evidence from the given labels of training nodes, the Graph-based Kernel Dirichlet distribution Estimation (GKDE) method is designed for accurately predicting node-level Dirichlet distributions and detecting out-of-distribution (OOD) nodes. We validated the outperformance of our proposed model compared to the state-of-the-art counterparts in terms of misclassification detection and OOD detection based on six real network datasets. We found that dissonance-based detection yielded the best results on misclassification detection while vacuity-based detection was the best for OOD detection. To clarify the reasons behind the results, we provided the theoretical proof that explains the relationships between different types of uncertainties considered in this work.

ConvBERT: Improving BERT with Span-based Dynamic Convolution

- Zi-Hang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, Shuicheng Yan
- abstract@[open-review](#): Pre-trained language models like BERT and its variants have recently achieved impressive performance in various natural language understanding tasks. However, BERT heavily relies on the global self-attention block and thus suffers large memory footprint and computation cost. Although all its attention heads query on the whole input sequence for generating the attention map from a global perspective, we observe some heads only need to learn local dependencies, which means existence of computation redundancy. We therefore propose a novel span-based dynamic convolution to replace these self-attention heads to directly model local dependencies. The novel convolution heads, together with the rest self-attention heads, form a new mixed attention block that is more efficient at both global and local context learning. We equip BERT with this mixed attention design and build a ConvBERT model. Experiments have shown that ConvBERT significantly outperforms BERT and its variants in various downstream tasks, with lower training cost and fewer model parameters. Remarkably, ConvBERTbase model achieves 86.4 GLUE score, 0.7 higher than ELECTRAbase, using less than 1/4 training cost. Code and pre-trained models will be released.

Practical No-box Adversarial Attacks against DNNs

- Qizhang Li, Yiwen Guo, Hao Chen
- abstract@[open-review](#): The study of adversarial vulnerabilities of deep neural networks (DNNs) has progressed rapidly. Existing attacks require either internal access (to the architecture, parameters, or training set of the victim model) or external access (to query the model). However, both the access may be infeasible or expensive in many scenarios. We investigate no-box adversarial examples, where the attacker can neither access the model information or the training set nor query the model. Instead, the attacker can only gather a small number of examples from the same problem domain as that of the victim model. Such a stronger threat model greatly expands the applicability of adversarial attacks. We propose three mechanisms for training with a very small dataset (on the order of tens of examples) and find that prototypical reconstruction is the most effective. Our experiments show that adversarial examples crafted on prototypical auto-encoding models transfer well to a variety of image classification and face verification models. On a commercial celebrity recognition system held by clarifai.com, our approach significantly diminishes the average prediction accuracy of the system to only 15.40%, which is on par with the attack that transfers adversarial examples from a pre-trained Arcface model. Our code is publicly available at: <https://github.com/qizhangli/nobox-attacks>.

Breaking the Sample Size Barrier in Model-Based Reinforcement Learning with a Generative Model

- Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, Yuxin Chen
- abstract@[open-review](#): We investigate the sample efficiency of reinforcement learning in a γ -discounted infinite-horizon Markov decision process (MDP) with state space S and action space A , assuming access to a generative model. Despite a number of prior work tackling this problem, a complete picture of the trade-offs between sample complexity and statistical accuracy is yet to be determined. In particular, prior results suffer from a sample size barrier, in the sense that their claimed statistical guarantees hold only when the sample size exceeds at least $|S||A| / (1-\gamma)^2$ (up to some log factor). The current paper overcomes this barrier by certifying the minimax optimality of model-based reinforcement learning as soon as the sample size exceeds the order of $|S||A| / (1-\gamma)$ (modulo some log factor). More specifically, a perturbed model-based planning algorithm provably finds an ϵ -optimal policy with an order of $|S||A| / ((1-\gamma)^3\epsilon^2)$ samples (up to log factor) for any $0 < \epsilon < 1/(1-\gamma)$. Along the way, we derive improved (instance-dependent) guarantees for model-based policy evaluation. To the best of our knowledge, this work provides the first minimax-optimal guarantee in a generative model that accommodates the entire range of sample sizes (beyond which finding a meaningful policy is information theoretically impossible).

Walking in the Shadow: A New Perspective on Descent Directions for Constrained Minimization

- Hassan Mortagy, Swati Gupta, Sebastian Pokutta
- abstract@[open-review](#): Descent directions such as movement towards Frank-Wolfe vertices, away steps, in-face away steps and pairwise directions have been an important design consideration in conditional gradient descent (CGD) variants. In this work, we attempt to demystify the impact of movement in these directions towards attaining constrained minimizers. The best local direction of descent is the directional derivative of the projection of the gradient, which we refer to as the "shadow" of the gradient. We show that the continuous-time dynamics of moving in the shadow are equivalent to those of PGD however non-trivial to discretize. By projecting gradients in PGD, one not only ensures feasibility but also is able to "wrap" around the convex region. We show that Frank-Wolfe (FW) vertices in fact recover the maximal wrap one can obtain by projecting gradients, thus providing a new perspective to these steps. We also claim that the shadow steps give the best direction of descent emanating from the convex hull of all possible away-vertices. Opening up the PGD movements in terms of shadow steps gives linear convergence, dependent on the number of faces. We combine these insights into a novel Shadow-CG method that uses FW steps (i.e., wrap around the polytope) and shadow steps (i.e., optimal local descent direction), while enjoying linear convergence. Our analysis develops properties of directional derivatives of projections (which may be of independent interest), while providing a unifying view of various descent directions in the CGD literature.

Path Sample-Analytic Gradient Estimators for Stochastic Binary Networks

- Alexander Shekhovtsov, Viktor Yanush, Boris Flach
- abstract@[open-review](#): In neural networks with binary activations and or binary weights the training by gradient descent is complicated as the model has piecewise constant response. We consider stochastic binary networks, obtained by adding noises in front of activations. The expected model response becomes a smooth function of parameters, its gradient is well defined but it is challenging to estimate it accurately. We propose a new method for this estimation problem combining sampling and analytic approximation steps. The method has a significantly reduced variance at the price of a small bias which gives a very practical tradeoff in comparison with existing unbiased and biased estimators. We further show that one extra linearization step leads to a deep straight-through estimator previously known only as an ad-hoc heuristic. We experimentally show higher accuracy in gradient estimation and demonstrate a more stable and better performing training in deep convolutional models with both proposed methods.

Reward Propagation Using Graph Convolutional Networks

- Martin Klissarov, Doina Precup
- abstract@[open-review](#): Potential-based reward shaping provides an approach for designing good reward functions, with the purpose of speeding up learning. However, automatically finding potential functions for complex environments is a difficult problem (in fact, of the same difficulty as learning a value function from scratch). We propose a new framework for learning potential functions by leveraging ideas from graph representation learning. Our approach relies on Graph Convolutional Networks which we use as a key ingredient in combination with the probabilistic inference view of reinforcement learning. More precisely, we leverage Graph Convolutional Networks to perform message passing from rewarding states. The propagated messages can then be used as potential functions for reward shaping to accelerate learning. We verify empirically that our approach can achieve considerable improvements in both small and high-dimensional control problems.

[LoopReg: Self-supervised Learning of Implicit Surface Correspondences, Pose and Shape for 3D Human Mesh Registration](#)

- Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, Gerard Pons-Moll
- abstract@[open-review](#): We address the problem of fitting 3D human models to 3D scans of dressed humans. Classical methods optimize both the data-to-model correspondences and the human model parameters (pose and shape), but are reliable only when initialised close to the solution. Some methods initialize the optimization based on fully supervised correspondence predictors, which is not differentiable end-to-end, and can only process a single scan at a time. Our main contribution is LoopReg, an end-to-end learning framework to register a corpus of scans to a common 3D human model. The key idea is to create a self-supervised loop. A backward map, parameterized by a Neural Network, predicts the correspondence from every scan point to the surface of the human model. A forward map, parameterized by a human model, transforms the corresponding points back to the scan based on the model parameters (pose and shape), thus closing the loop. Formulating this closed loop is not straightforward because it is not trivial to force the output of the NN to be on the surface of the human model -- outside this surface the human model is not even defined. To this end, we propose two key innovations. First, we define the canonical surface implicitly as the zero level set of a distance field in R3, which in contrast to more common UV parameterizations does not require cutting the surface, does not have discontinuities, and does not induce distortion. Second, we diffuse the human model to the 3D domain. This allows to map the NN predictions forward, even when they slightly deviate from the zero level set. Results demonstrate that we can train LoopReg mainly self-supervised -- following a supervised warm-start, the model becomes increasingly more accurate as additional unlabelled raw scans are processed. Our code and pre-trained models can be downloaded for research.

[Fully Dynamic Algorithm for Constrained Submodular Optimization](#)

- Silvio Lattanzi, Slobodan Mitrović, Ashkan Norouzi-Fard, Jakub M. Tarnawski, Morteza Zadimoghaddam
- abstract@[open-review](#): The task of maximizing a monotone submodular function under a cardinality constraint is at the core of many machine learning and data mining applications, including data summarization, sparse regression and coverage problems. We study this classic problem in the fully dynamic setting, where elements can be both inserted and removed. Our main result is a randomized algorithm that maintains an efficient data structure with a polylogarithmic amortized update time and yields a $(1/2 - \epsilon)$ -approximate solution. We complement our theoretical analysis with an empirical study of the performance of our algorithm.

[Robust Optimal Transport with Applications in Generative Modeling and Domain Adaptation](#)

- Yogesh Balaji, Rama Chellappa, Soheil Feizi
- abstract@[open-review](#): Optimal Transport (OT) distances such as Wasserstein have been used in several areas such as GANs and domain adaptation. OT, however, is very sensitive to outliers (samples with large noise) in the data since in its objective function, every sample, including outliers, is weighed similarly due to the marginal constraints. To remedy this issue, robust formulations of OT with unbalanced marginal constraints have previously been proposed. However, employing these methods in deep learning problems such as GANs and domain adaptation is challenging due to the instability of their dual optimization solvers. In this paper, we resolve these issues by deriving a computationally-efficient dual form of the robust OT optimization that is amenable to modern deep learning applications. We demonstrate the effectiveness of our formulation in two applications of GANs and domain adaptation. Our approach can train state-of-the-art GAN models on noisy datasets corrupted with outlier distributions. In particular, the proposed optimization method computes weights for training samples reflecting how difficult it is for those samples to be generated in the model. In domain adaptation, our robust OT formulation leads to improved accuracy compared to the standard adversarial adaptation methods. Our code is available at <https://github.com/yogeshbalaji/robustOT>.

[Autofocused oracles for model-based design](#)

- Clara Fannjiang, Jennifer Listgarten
- abstract@[open-review](#): Data-driven design is making headway into a number of application areas, including protein, small-molecule, and materials engineering. The design goal is to construct an object with desired properties, such as a protein that binds to a therapeutic target, or a superconducting material with a higher critical temperature than previously observed. To that end, costly experimental measurements are being replaced with calls to high-capacity regression models trained on labeled data, which can be leveraged in an in silico search for design candidates. However, the design goal necessitates moving into regions of the design space beyond where such models were trained. Therefore, one can ask: should the regression model be altered as the design algorithm explores the design space, in the absence of new data? Herein, we answer this question in the affirmative. In particular, we (i) formalize the data-driven design problem as a non-zero-sum game, (ii) develop a principled strategy for retraining the regression model as the design algorithm proceeds---what we refer to as autofocusing, and (iii) demonstrate the promise of autofocusing empirically.

[Debiasing Averaged Stochastic Gradient Descent to handle missing values](#)

- Aude Sportisse, Claire Boyer, Aymeric Dieuleveut, Julie Josse
- abstract@[open-review](#): Stochastic gradient algorithm is a key ingredient of many machine learning methods, particularly appropriate for large-scale learning. However, a major caveat of large data is their incompleteness. We propose an averaged stochastic gradient algorithm handling missing values in linear models. This approach has the merit to be free from the need of any data distribution modeling and to account for heterogeneous missing proportion. In both streaming and finite-sample settings, we prove that this algorithm achieves convergence rate of $\mathcal{O}(\frac{1}{n})$ at the iteration n , the same as without missing values. We show the convergence behavior and the relevance of the algorithm not only on synthetic data but also on real data sets, including those collected from medical register.

[Trajectory-wise Multiple Choice Learning for Dynamics Generalization in Reinforcement Learning](#)

- Younggyo Seo, Kimin Lee, Ignasi Clavera Gilaberte, Thanard Kurutach, Jinwoo Shin, Pieter Abbeel
- abstract@[open-review](#): Model-based reinforcement learning (RL) has shown great potential in various control tasks in terms of both sample-efficiency and final performance. However, learning a generalizable dynamics model robust to changes in dynamics remains a challenge since the target transition dynamics follow a multi-modal distribution. In this paper, we present a new model-based RL algorithm, coined trajectory-wise multiple choice learning, that learns a multi-headed dynamics model for dynamics generalization. The main idea is updating the most accurate prediction head to specialize each head in certain environments with similar dynamics, i.e., clustering environments. Moreover, we incorporate context learning, which encodes dynamics-specific information from past experiences into the context latent vector, enabling the model to perform online adaptation to unseen environments. Finally, to utilize the specialized prediction heads more effectively, we propose an adaptive planning method, which selects the most accurate prediction head over

a recent experience. Our method exhibits superior zero-shot generalization performance across a variety of control tasks, compared to state-of-the-art RL methods. Source code and videos are available at <https://sites.google.com/view/trajectory-mcl>.

[CompRess: Self-Supervised Learning by Compressing Representations](#)

- Soroush Abbasi Koohpayegani, Ajinkya Tejankar, Hamed Pirsiavash
- abstract@[open-review](#): Self-supervised learning aims to learn good representations with unlabeled data. Recent works have shown that larger models benefit more from self-supervised learning than smaller models. As a result, the gap between supervised and self-supervised learning has been greatly reduced for larger models. In this work, instead of designing a new pseudo task for self-supervised learning, we develop a model compression method to compress an already learned, deep self-supervised model (teacher) to a smaller one (student). We train the student model so that it mimics the relative similarity between the datapoints in the teacher's embedding space. For AlexNet, our method outperforms all previous methods including the fully supervised model on ImageNet linear evaluation (59.0% compared to 56.5%) and on nearest neighbor evaluation (50.7% compared to 41.4%). To the best of our knowledge, this is the first time a self-supervised AlexNet has outperformed supervised one on ImageNet classification. Our code is available here: <https://github.com/UMBCvision/CompRess>

[Sample complexity and effective dimension for regression on manifolds](#)

- Andrew McRae, Justin Romberg, Mark Davenport
- abstract@[open-review](#): We consider the theory of regression on a manifold using reproducing kernel Hilbert space methods. Manifold models arise in a wide variety of modern machine learning problems, and our goal is to help understand the effectiveness of various implicit and explicit dimensionality-reduction methods that exploit manifold structure. Our first key contribution is to establish a novel nonasymptotic version of the Weyl law from differential geometry. From this we are able to show that certain spaces of smooth functions on a manifold are effectively finite-dimensional, with a complexity that scales according to the manifold dimension rather than any ambient data dimension. Finally, we show that given (potentially noisy) function values taken uniformly at random over a manifold, a kernel regression estimator (derived from the spectral decomposition of the manifold) yields minimax-optimal error bounds that are controlled by the effective dimension.

[The phase diagram of approximation rates for deep neural networks](#)

- Dmitry Yarotsky, Anton Zhevnerchuk
- abstract@[open-review](#): We explore the phase diagram of approximation rates for deep neural networks and prove several new theoretical results. In particular, we generalize the existing result on the existence of deep discontinuous phase in ReLU networks to functional classes of arbitrary positive smoothness, and identify the boundary between the feasible and infeasible rates. Moreover, we show that all networks with a piecewise polynomial activation function have the same phase diagram. Next, we demonstrate that standard fully-connected architectures with a fixed width independent of smoothness can adapt to smoothness and achieve almost optimal rates. Finally, we consider deep networks with periodic activations ("deep Fourier expansion") and prove that they have very fast, nearly exponential approximation rates, thanks to the emerging capability of the network to implement efficient lookup operations.

[Timeseries Anomaly Detection using Temporal Hierarchical One-Class Network](#)

- Lifeng Shen, Zhuocong Li, James Kwok
- abstract@[open-review](#): Real-world timeseries have complex underlying temporal dynamics and the detection of anomalies is challenging. In this paper, we propose the Temporal Hierarchical One-Class (THOC) network, a temporal one-class classification model for timeseries anomaly detection. It captures temporal dynamics in multiple scales by using a dilated recurrent neural network with skip connections. Using multiple hyperspheres obtained with a hierarchical clustering process, a one-class objective called Multiscale Vector Data Description is defined. This allows the temporal dynamics to be well captured by a set of multi-resolution temporal clusters. To further facilitate representation learning, the hypersphere centers are encouraged to be orthogonal to each other, and a self-supervision task in the temporal domain is added. The whole model can be trained end-to-end. Extensive empirical studies on various real-world timeseries demonstrate that the proposed THOC network outperforms recent strong deep learning baselines on timeseries anomaly detection.

[EcoLight: Intersection Control in Developing Regions Under Extreme Budget and Network Constraints](#)

- Sachin Chauhan, Kashish Bansal, Rijurekha Sen
- abstract@[open-review](#): Effective intersection control can play an important role in reducing traffic congestion and associated vehicular emissions. This is vitally needed in developing countries, where air pollution is reaching life threatening levels. This paper presents EcoLight intersection control for developing regions, where budget is constrained and network connectivity is very poor. EcoLight learns effective control offline using state-of-the-art Deep Reinforcement Learning methods, but deploys highly efficient runtime control algorithms on low cost embedded devices that work stand-alone on road without server connectivity. EcoLight optimizes both average case and worst case values of throughput, travel time and other metrics, as evaluated on open-source datasets from New York and on a custom developing region dataset.

[Reconstructing Perceptive Images from Brain Activity by Shape-Semantic GAN](#)

- Tao Fang, Yu Qi, Gang Pan
- abstract@[open-review](#): Reconstructing seeing images from fMRI recordings is an absorbing research area in neuroscience and provides a potential brain-reading technology. The challenge lies in that visual encoding in brain is highly complex and not fully revealed. Inspired by the theory that visual features are hierarchically represented in cortex, we propose to break the complex visual signals into multi-level components and decode each component separately. Specifically, we decode shape and semantic representations from the lower and higher visual cortex respectively, and merge the shape and semantic information to images by a generative adversarial network (Shape-Semantic GAN). This 'divide and conquer' strategy captures visual information more accurately. Experiments demonstrate that Shape-Semantic GAN improves the reconstruction similarity and image quality, and achieves the state-of-the-art image reconstruction performance.

[Emergent Complexity and Zero-shot Transfer via Unsupervised Environment Design](#)

- Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell, Andrew Critch, Sergey Levine
- abstract@[open-review](#): A wide range of reinforcement learning (RL) problems --- including robustness, transfer learning, unsupervised RL, and emergent complexity --- require specifying a distribution of tasks or environments in which a policy will be trained. However, creating a useful distribution of environments is error prone, and takes a significant amount of developer time and effort. We propose Unsupervised Environment Design (UED) as an alternative paradigm, where developers provide environments with unknown parameters, and these parameters are used to automatically produce a distribution over valid, solvable environments. Existing approaches to automatically generating environments suffer from common failure modes: domain randomization cannot generate structure or adapt the difficulty of the environment to the agent's learning progress, and minimax adversarial training leads to worst-case environments that are often unsolvable. To generate structured, solvable environments for our protagonist agent, we introduce a second, antagonist agent that is allied with the environment-generating adversary. The adversary is motivated to generate environments which maximize regret,

defined as the difference between the protagonist and antagonist agent's return. We call our technique Protagonist Antagonist Induced Regret Environment Design (PAIRED). Our experiments demonstrate that PAIRED produces a natural curriculum of increasingly complex environments, and PAIRED agents achieve higher zero-shot transfer performance when tested in highly novel environments.

[A Spectral Energy Distance for Parallel Speech Synthesis](#)

- Alexey Gritsenko, Tim Salimans, Rianne van den Berg, Jasper Snoek, Nal Kalchbrenner
- abstract@[open-review](#): Speech synthesis is an important practical generative modeling problem that has seen great progress over the last few years, with likelihood-based autoregressive neural models now outperforming traditional concatenative systems. A downside of such autoregressive models is that they require executing tens of thousands of sequential operations per second of generated audio, making them ill-suited for deployment on specialized deep learning hardware. Here, we propose a new learning method that allows us to train highly parallel models of speech, without requiring access to an analytical likelihood function. Our approach is based on a generalized energy distance between the distributions of the generated and real audio. This spectral energy distance is a proper scoring rule with respect to the distribution over magnitude-spectrograms of the generated waveform audio and offers statistical consistency guarantees. The distance can be calculated from minibatches without bias, and does not involve adversarial learning, yielding a stable and consistent method for training implicit generative models. Empirically, we achieve state-of-the-art generation quality among implicit generative models, as judged by the recently-proposed cFDSD metric. When combining our method with adversarial techniques, we also improve upon the recently-proposed GAN-TTS model in terms of Mean Opinion Score as judged by trained human evaluators.

[Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations](#)

- Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, James J. DiCarlo
- abstract@[open-review](#): Current state-of-the-art object recognition models are largely based on convolutional neural network (CNN) architectures, which are loosely inspired by the primate visual system. However, these CNNs can be fooled by imperceptibly small, explicitly crafted perturbations, and struggle to recognize objects in corrupted images that are easily recognized by humans. Here, by making comparisons with primate neural data, we first observed that CNN models with a neural hidden layer that better matches primate primary visual cortex (V1) are also more robust to adversarial attacks. Inspired by this observation, we developed VOneNets, a new class of hybrid CNN vision models. Each VOneNet contains a fixed weight neural network front-end that simulates primate V1, called the VOneBlock, followed by a neural network back-end adapted from current CNN vision models. The VOneBlock is based on a classical neuroscientific model of V1: the linear-nonlinear-Poisson model, consisting of a biologically-constrained Gabor filter bank, simple and complex cell nonlinearities, and a V1 neuronal stochasticity generator. After training, VOneNets retain high ImageNet performance, but each is substantially more robust, outperforming the base CNNs and state-of-the-art methods by 18% and 3%, respectively, on a conglomerate benchmark of perturbations comprised of white box adversarial attacks and common image corruptions. Finally, we show that all components of the VOneBlock work in synergy to improve robustness. While current CNN architectures are arguably brain-inspired, the results presented here demonstrate that more precisely mimicking just one stage of the primate visual system leads to new gains in ImageNet-level computer vision applications.

[Learning from Positive and Unlabeled Data with Arbitrary Positive Shift](#)

- Zayd Hammoudeh, Daniel Lowd
- abstract@[open-review](#): Positive-unlabeled (PU) learning trains a binary classifier using only positive and unlabeled data. A common simplifying assumption is that the positive data is representative of the target positive class. This assumption rarely holds in practice due to temporal drift, domain shift, and/or adversarial manipulation. This paper shows that PU learning is possible even with arbitrarily non-representative positive data given unlabeled data from the source and target distributions. Our key insight is that only the negative class's distribution need be fixed. We integrate this into two statistically consistent methods to address arbitrary positive bias - one approach combines negative-unlabeled learning with unlabeled-unlabeled learning while the other uses a novel, recursive risk estimator. Experimental results demonstrate our methods' effectiveness across numerous real-world datasets and forms of positive bias, including disjoint positive class-conditional supports. Additionally, we propose a general, simplified approach to address PU risk estimation overfitting.

[Deep Energy-based Modeling of Discrete-Time Physics](#)

- Takashi Matsubara, Ai Ishikawa, Takaharu Yaguchi
- abstract@[open-review](#): Physical phenomena in the real world are often described by energy-based modeling theories, such as Hamiltonian mechanics or the Landau theory, which yield various physical laws. Recent developments in neural networks have enabled the mimicking of the energy conservation law by learning the underlying continuous-time differential equations. However, this may not be possible in discrete time, which is often the case in practical learning and computation. Moreover, other physical laws have been overlooked in the previous neural network models. In this study, we propose a deep energy-based physical model that admits a specific differential geometric structure. From this structure, the conservation or dissipation law of energy and the mass conservation law follow naturally. To ensure the energetic behavior in discrete time, we also propose an automatic discrete differentiation algorithm that enables neural networks to employ the discrete gradient method.

[Quantifying Learnability and Describability of Visual Concepts Emerging in Representation Learning](#)

- Iro Laina, Ruth Fong, Andrea Vedaldi
- abstract@[open-review](#): The increasing impact of black box models, and particularly of unsupervised ones, comes with an increasing interest in tools to understand and interpret them. In this paper, we consider in particular how to characterise visual groupings discovered automatically by deep neural networks, starting with state-of-the-art clustering methods. In some cases, clusters readily correspond to an existing labelled dataset. However, often they do not, yet they still maintain an "intuitive interpretability". We introduce two concepts, visual learnability and describability, that can be used to quantify the interpretability of arbitrary image groupings, including unsupervised ones. The idea is to measure (1) how well humans can learn to reproduce a grouping by measuring their ability to generalise from a small set of visual examples (learnability) and (2) whether the set of visual examples can be replaced by a succinct, textual description (describability). By assessing human annotators as classifiers, we remove the subjective quality of existing evaluation metrics. For better scalability, we finally propose a class-level captioning system to generate descriptions for visual groupings automatically and compare it to human annotators using the describability metric.

[Self-Learning Transformations for Improving Gaze and Head Redirection](#)

- Yufeng Zheng, Seonwook Park, Xucong Zhang, Shalini De Mello, Otmar Hilliges
- abstract@[open-review](#): Many computer vision tasks rely on labeled data. Rapid progress in generative modeling has led to the ability to synthesize photorealistic images. However, controlling specific aspects of the generation process such that the data can be used for supervision of downstream tasks remains challenging. In this paper we propose a novel generative model for images of faces, that is capable of producing high-quality images under fine-grained control over eye gaze and head orientation angles. This requires the disentangling of many appearance related factors including gaze and head orientation but also lighting, hue etc. We propose a novel architecture which learns to discover, disentangle and encode these extraneous variations in a self-learned manner. We further show that explicitly disentangling task-irrelevant factors results in more accurate modelling of gaze and head orientation. A novel evaluation scheme shows that our method improves upon the state-of-the-art in redirection accuracy and disentanglement between gaze direction and head orientation changes. Furthermore, we show that in the presence of limited amounts of real-world training data, our method allows for

improvements in the downstream task of semi-supervised cross-dataset gaze estimation. Please check our project page at:
<https://ait.ethz.ch/projects/2020/STED-gaze/>

[Language-Conditioned Imitation Learning for Robot Manipulation Tasks](#)

- Simon Stepputtis, Joseph Campbell, Mariano Phiellip, Stefan Lee, Chitta Baral, Heni Ben Amor
- abstract@[open-review](#): Imitation learning is a popular approach for teaching motor skills to robots. However, most approaches focus on extracting policy parameters from execution traces alone (i.e., motion trajectories and perceptual data). No adequate communication channel exists between the human expert and the robot to describe critical aspects of the task, such as the properties of the target object or the intended shape of the motion. Motivated by insights into the human teaching process, we introduce a method for incorporating unstructured natural language into imitation learning. At training time, the expert can provide demonstrations along with verbal descriptions in order to describe the underlying intent (e.g., "go to the large green bowl"). The training process then interrelates these two modalities to encode the correlations between language, perception, and motion. The resulting language-conditioned visuomotor policies can be conditioned at runtime on new human commands and instructions, which allows for more fine-grained control over the trained policies while also reducing situational ambiguity. We demonstrate in a set of simulation experiments how our approach can learn language-conditioned manipulation policies for a seven-degree-of-freedom robot arm and compare the results to a variety of alternative methods.

[POMDPs in Continuous Time and Discrete Spaces](#)

- Bastian Alt, Matthias Schultheis, Heinz Koeppl
- abstract@[open-review](#): Many processes, such as discrete event systems in engineering or population dynamics in biology, evolve in discrete space and continuous time. We consider the problem of optimal decision making in such discrete state and action space systems under partial observability. This places our work at the intersection of optimal filtering and optimal control. At the current state of research, a mathematical description for simultaneous decision making and filtering in continuous time with finite state and action spaces is still missing. In this paper, we give a mathematical description of a continuous-time partial observable Markov decision process (POMDP). By leveraging optimal filtering theory we derive a Hamilton-Jacobi-Bellman (HJB) type equation that characterizes the optimal solution. Using techniques from deep learning we approximately solve the resulting partial integro-differential equation. We present (i) an approach solving the decision problem offline by learning an approximation of the value function and (ii) an online algorithm which provides a solution in belief space using deep reinforcement learning. We show the applicability on a set of toy examples which pave the way for future methods providing solutions for high dimensional problems.

[Exemplar Guided Active Learning](#)

- Jason S. Hartford, Kevin Leyton-Brown, Hadas Raviv, Dan Padnos, Shahar Lev, Barak Lenz
- abstract@[open-review](#): We consider the problem of wisely using a limited budget to label a small subset of a large unlabeled dataset. For example, consider the NLP problem of word sense disambiguation. For any word, we have a set of candidate labels from a knowledge base, but the label set is not necessarily representative of what occurs in the data: there may exist labels in the knowledge base that very rarely occur in the corpus because the sense is rare in modern English; and conversely there may exist true labels that do not exist in our knowledge base. Our aim is to obtain a classifier that performs as well as possible on examples of each â€œcommon classâ€ that occurs with frequency above a given threshold in the unlabeled set while annotating as few examples as possible from â€œrare classesâ€ whose labels occur with less than this frequency. The challenge is that we are not informed which labels are common and which are rare, and the true label distribution may exhibit extreme skew. We describe an active learning approach that (1) explicitly searches for rare classes by leveraging the contextual embedding spaces provided by modern language models, and (2) incorporates a stopping rule that ignores classes once we prove that they occur below our target threshold with high probability. We prove that our algorithm only costs logarithmically more than a hypothetical approach that knows all true label frequencies and show experimentally that incorporating automated search can significantly reduce the number of samples needed to reach target accuracy levels.

[Grasp Proposal Networks: An End-to-End Solution for Visual Learning of Robotic Grasps](#)

- Chaozheng Wu, Jian Chen, Qiaoyu Cao, Jianchi Zhang, Yunxin Tai, Lin Sun, Kui Jia
- abstract@[open-review](#): Learning robotic grasps from visual observations is a promising yet challenging task. Recent research shows its great potential by preparing and learning from large-scale synthetic datasets. For the popular, 6 degree-of-freedom (6-DOF) grasp setting of parallel-jaw gripper, most of existing methods take the strategy of heuristically sampling grasp candidates and then evaluating them using learned scoring functions. This strategy is limited in terms of the conflict between sampling efficiency and coverage of optimal grasps. To this end, we propose in this work a novel, end-to-end \emph{Grasp Proposal Network (GPNet)}, to predict a diverse set of 6-DOF grasps for an unseen object observed from a single and unknown camera view. GPNet builds on a key design of grasp proposal module that defines \emph{anchors of grasp centers} at discrete but regular 3D grid corners, which is flexible to support either more precise or more diverse grasp predictions. To test GPNet, we contribute a synthetic dataset of 6-DOF object grasps; evaluation is conducted using rule-based criteria, simulation test, and real test. Comparative results show the advantage of our methods over existing ones. Notably, GPNet gains better simulation results via the specified coverage, which helps achieve a ready translation in real test. Our code and dataset are available on \url{https://github.com/CZ-Wu/GPNet}.

[Node Embeddings and Exact Low-Rank Representations of Complex Networks](#)

- Sudhanshu Chanpuriya, Cameron Musco, Konstantinos Sotiropoulos, Charalampos Tsourakakis
- abstract@[open-review](#): Low-dimensional embeddings, from classical spectral embeddings to modern neural-net-inspired methods, are a cornerstone in the modeling and analysis of complex networks. Recent work by Seshadhri et al. (PNAS 2020) suggests that such embeddings cannot capture local structure arising in complex networks. In particular, they show that any network generated from a natural low-dimensional model cannot be both sparse and have high triangle density (high clustering coefficient), two hallmark properties of many real-world networks. In this work we show that the results of Seshadhri et al. are intimately connected to the model they use rather than the low-dimensional structure of complex networks. Specifically, we prove that a minor relaxation of their model can generate sparse graphs with high triangle density. Surprisingly, we show that this same model leads to exact low-dimensional factorizations of many real-world networks. We give a simple algorithm based on logistic principal component analysis (LPCA) that succeeds in finding such exact embeddings. Finally, we perform a large number of experiments that verify the ability of very low-dimensional embeddings to capture local structure in real-world networks.

[Fictitious Play for Mean Field Games: Continuous Time Analysis and Applications](#)

- Sarah Perrin, Julien Perolat, Mathieu Lauriere, Matthieu Geist, Romuald Elie, Olivier Pietquin
- abstract@[open-review](#): In this paper, we deepen the analysis of continuous time Fictitious Play learning algorithm to the consideration of various finite state Mean Field Game settings (finite horizon, $\$\\gamma\$$ -discounted), allowing in particular for the introduction of an additional common noise. We first present a theoretical convergence analysis of the continuous time Fictitious Play process and prove that the induced exploitability decreases at a rate $O(\\frac{1}{t})$. Such analysis emphasizes the use of exploitability as a relevant metric for evaluating the convergence towards a Nash equilibrium in the context of Mean Field Games. These theoretical contributions are supported by numerical experiments provided in either model-based or model-free settings. We provide hereby for the first time converging learning dynamics for Mean Field Games in the presence of common noise.

Steering Distortions to Preserve Classes and Neighbors in Supervised Dimensionality Reduction

- Benoît Colange, Jaakko Peltonen, Michael Aupetit, Denys Dutykh, Sylvain Lespinats
- abstract@[open-review](#): Nonlinear dimensionality reduction of high-dimensional data is challenging as the low-dimensional embedding will necessarily contain distortions, and it can be hard to determine which distortions are the most important to avoid. When annotation of data into known relevant classes is available, it can be used to guide the embedding to avoid distortions that worsen class separation. The supervised mapping method introduced in the present paper, called ClassNeRV, proposes an original stress function that takes class annotation into account and evaluates embedding quality both in terms of false neighbors and missed neighbors. ClassNeRV shares the theoretical framework of a family of methods descended from Stochastic Neighbor Embedding (SNE). Our approach has a key advantage over previous ones: in the literature supervised methods often emphasize class separation at the price of distorting the data neighbors' structure; conversely, unsupervised methods provide better preservation of structure at the price of often mixing classes. Experiments show that ClassNeRV can preserve both neighbor structure and class separation, outperforming nine state of the art alternatives.

On Infinite-Width Hypernetworks

- Eta Littwin, Tomer Galanti, Lior Wolf, Greg Yang
- abstract@[open-review](#): {em Hypernetworks} are architectures that produce the weights of a task-specific {em primary network}. A notable application of hypernetworks in the recent literature involves learning to output functional representations. In these scenarios, the hypernetwork learns a representation corresponding to the weights of a shallow MLP, which typically encodes shape or image information. While such representations have seen considerable success in practice, they remain lacking in the theoretical guarantees in the wide regime of the standard architectures. In this work, we study wide over-parameterized hypernetworks. We show that unlike typical architectures, infinitely wide hypernetworks do not guarantee convergence to a global minima under gradient descent. We further show that convexity can be achieved by increasing the dimensionality of the hypernetwork's output, to represent wide MLPs. In the dually infinite-width regime, we identify the functional priors of these architectures by deriving their corresponding GP and NTK kernels, the latter of which we refer to as the {em hyperkernel}. As part of this study, we make a mathematical contribution by deriving tight bounds on high order Taylor expansion terms of standard fully connected ReLU networks.

Interferobot: aligning an optical interferometer by a reinforcement learning agent

- Dmitry Sorokin, Alexander Ulanov, Ekaterina Sazhina, Alexander Lvovsky
- abstract@[open-review](#): Limitations in acquiring training data restrict potential applications of deep reinforcement learning (RL) methods to the training of real-world robots. Here we train an RL agent to align a Mach-Zehnder interferometer, which is an essential part of many optical experiments, based on images of interference fringes acquired by a monocular camera. The agent is trained in a simulated environment, without any hand-coded features or a priori information about the physics, and subsequently transferred to a physical interferometer. Thanks to a set of domain randomizations simulating uncertainties in physical measurements, the agent successfully aligns this interferometer without any fine-tuning, achieving a performance level of a human expert.

Program Synthesis with Pragmatic Communication

- Yewen Pu, Kevin Ellis, Marta Kryven, Josh Tenenbaum, Armando Solar-Lezama
- abstract@[open-review](#): Program synthesis techniques construct or infer programs from user-provided specifications, such as input-output examples. Yet most specifications, especially those given by end-users, leave the synthesis problem radically ill-posed, because many programs may simultaneously satisfy the specification. Prior work resolves this ambiguity by using various inductive biases, such as a preference for simpler programs. This work introduces a new inductive bias derived by modeling the program synthesis task as rational communication, drawing insights from recursive reasoning models of pragmatics. Given a specification, we score a candidate program both on its consistency with the specification, and also whether a rational speaker would choose this particular specification to communicate that program. We develop efficient algorithms for such an approach when learning from input-output examples, and build a pragmatic program synthesizer over a simple grid-like layout domain. A user study finds that end-user participants communicate more effectively with the pragmatic program synthesizer over a non-pragmatic one.

Principal Neighbourhood Aggregation for Graph Nets

- Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, Petar Veličković
- abstract@[open-review](#): Graph Neural Networks (GNNs) have been shown to be effective models for different predictive tasks on graph-structured data. Recent work on their expressive power has focused on isomorphism tasks and countable feature spaces. We extend this theoretical framework to include continuous features---which occur regularly in real-world input domains and within the hidden layers of GNNs---and we demonstrate the requirement for multiple aggregation functions in this context. Accordingly, we propose Principal Neighbourhood Aggregation (PNA), a novel architecture combining multiple aggregators with degree-scalers (which generalize the sum aggregator). Finally, we compare the capacity of different models to capture and exploit the graph structure via a novel benchmark containing multiple tasks taken from classical graph theory, alongside existing benchmarks from real-world domains, all of which demonstrate the strength of our model. With this work we hope to steer some of the GNN research towards new aggregation methods which we believe are essential in the search for powerful and robust models.

Reliable Graph Neural Networks via Robust Aggregation

- Simon Geisler, Daniel Zügner, Stephan Günnemann
- abstract@[open-review](#): Perturbations targeting the graph structure have proven to be extremely effective in reducing the performance of Graph Neural Networks (GNNs), and traditional defenses such as adversarial training do not seem to be able to improve robustness. This work is motivated by the observation that adversarially injected edges effectively can be viewed as additional samples to a node's neighborhood aggregation function, which results in distorted aggregations accumulating over the layers. Conventional GNN aggregation functions, such as a sum or mean, can be distorted arbitrarily by a single outlier. We propose a robust aggregation function motivated by the field of robust statistics. Our approach exhibits the largest possible breakdown point of 0.5, which means that the bias of the aggregation is bounded as long as the fraction of adversarial edges of a node is less than 50%. Our novel aggregation function, Soft Medoid, is a fully differentiable generalization of the Medoid and therefore lends itself well for end-to-end deep learning. Equipping a GNN with our aggregation improves the robustness with respect to structure perturbations on Cora ML by a factor of 3 (and 5.5 on Citeseer) and by a factor of 8 for low-degree nodes.

Instance Selection for GANs

- Terrance DeVries, Michal Drozdzal, Graham W. Taylor
- abstract@[open-review](#): Recent advances in Generative Adversarial Networks (GANs) have led to their widespread adoption for the purposes of generating high quality synthetic imagery. While capable of generating photo-realistic images, these models often produce unrealistic samples which fall outside of the data manifold. Several recently proposed techniques attempt to avoid spurious samples, either by rejecting them after generation, or by truncating the model's latent space. While effective, these methods are inefficient, as a large fraction of training time and model capacity are dedicated towards samples that will ultimately go unused. In this work we propose a novel approach to improve sample quality: altering the training dataset via instance selection before model training has taken place. By refining the empirical data distribution before training, we redirect model capacity towards high-density regions,

which ultimately improves sample fidelity, lowers model capacity requirements, and significantly reduces training time. Code is available at https://github.com/uoguelph-mlrg/instanceselectionfor_gans.

[Linear Disentangled Representations and Unsupervised Action Estimation](#)

- Matthew Painter, Adam Prugel-Bennett, Jonathon Hare
- abstract@[open-review](#): Disentangled representation learning has seen a surge in interest over recent times, generally focusing on new models which optimise one of many disparate disentanglement metrics. Symmetry Based Disentangled Representation learning introduced a robust mathematical framework that defined precisely what is meant by a ``linear disentangled representation''. This framework determined that such representations would depend on a particular decomposition of the symmetry group acting on the data, showing that actions would manifest through irreducible group representations acting on independent representational subspaces. \citet{forwardvae} subsequently proposed the first model to induce and demonstrate a linear disentangled representation in a VAE model. In this work we empirically show that linear disentangled representations are not present in standard VAE models and that they instead require altering the loss landscape to induce them. We proceed to show that such representations are a desirable property with regard to classical disentanglement metrics. Finally we propose a method to induce irreducible representations which forgoes the need for labelled action sequences, as was required by prior work. We explore a number of properties of this method, including the ability to learn from action sequences without knowledge of intermediate states and robustness under visual noise. We also demonstrate that it can successfully learn 4 independent symmetries directly from pixels.

[Video Frame Interpolation without Temporal Priors](#)

- Youjian Zhang, Chaoyue Wang, Dacheng Tao
- abstract@[open-review](#): Video frame interpolation, which aims to synthesize non-exist intermediate frames in a video sequence, is an important research topic in computer vision. Existing video frame interpolation methods have achieved remarkable results under specific assumptions, such as instant or known exposure time. However, in complicated real-world situations, the temporal priors of videos, i.e. frames per second (FPS) and frame exposure time, may vary from different camera sensors. When test videos are taken under different exposure settings from training ones, the interpolated frames will suffer significant misalignment problems. In this work, we solve the video frame interpolation problem in a general situation, where input frames can be acquired under uncertain exposure (and interval) time. Unlike previous methods that can only be applied to a specific temporal prior, we derive a general curvilinear motion trajectory formula from four consecutive sharp frames or two consecutive blurry frames without temporal priors. Moreover, utilizing constraints within adjacent motion trajectories, we devise a novel optical flow refinement strategy for better interpolation results. Finally, experiments demonstrate that one well-trained model is enough for synthesizing high-quality slow-motion videos under complicated real-world situations. Codes are available on <https://github.com/yjzhang96/UTI-VFI>.

[Learning compositional functions via multiplicative weight updates](#)

- Jeremy Bernstein, Jiawei Zhao, Markus Meister, Ming-Yu Liu, Anima Anandkumar, Yisong Yue
- abstract@[open-review](#): Compositionality is a basic structural feature of both biological and artificial neural networks. Learning compositional functions via gradient descent incurs well known problems like vanishing and exploding gradients, making careful learning rate tuning essential for real-world applications. This paper proves that multiplicative weight updates satisfy a descent lemma tailored to compositional functions. Based on this lemma, we derive Madam---a multiplicative version of the Adam optimiser---and show that it can train state of the art neural network architectures without learning rate tuning. We further show that Madam is easily adapted to train natively compressed neural networks by representing their weights in a logarithmic number system. We conclude by drawing connections between multiplicative weight updates and recent findings about synapses in biology.

[Sample Complexity of Uniform Convergence for Multicalibration](#)

- Eliran Shabat, Lee Cohen, Yishay Mansour
- abstract@[open-review](#): There is a growing interest in societal concerns in machine learning systems, especially in fairness. Multicalibration gives a comprehensive methodology to address group fairness. In this work, we address the multicalibration error and decouple it from the prediction error. The importance of decoupling the fairness metric (multicalibration) and the accuracy (prediction error) is due to the inherent trade-off between the two, and the societal decision regarding the ``right tradeoff'' (as imposed many times by regulators). Our work gives sample complexity bounds for uniform convergence guarantees of multicalibration error, which implies that regardless of the accuracy, we can guarantee that the empirical and (true) multicalibration errors are close. We emphasize that our results: (1) are more general than previous bounds, as they apply to both agnostic and realizable settings, and do not rely on a specific type of algorithm (such as differentially private), (2) improve over previous multicalibration sample complexity bounds and (3) implies uniform convergence guarantees for the classical calibration error.

[Differentiable Neural Architecture Search in Equivalent Space with Exploration Enhancement](#)

- Miao Zhang, Huiqi Li, Shirui Pan, Xiaojun Chang, Zongyuan Ge, Steven Su
- abstract@[open-review](#): Recent works on One-Shot Neural Architecture Search (NAS) mostly adopt a bilevel optimization scheme to alternatively optimize the supernet weights and architecture parameters after relaxing the discrete search space into a differentiable space. However, the non-negligible incongruence in their relaxation methods is hard to guarantee the differentiable optimization in the continuous space is equivalent to the optimization in the discrete space. Differently, this paper utilizes a variational graph autoencoder to injectively transform the discrete architecture space into an equivalently continuous latent space, to resolve the incongruence. A probabilistic exploration enhancement method is accordingly devised to encourage intelligent exploration during the architecture search in the latent space, to avoid local optimal in architecture search. As the catastrophic forgetting in differentiable One-Shot NAS deteriorates supernet predictive ability and makes the bilevel optimization inefficient, this paper further proposes an architecture complementation method to relieve this deficiency. We analyze the effectiveness of the proposed method, and a series of experiments have been conducted to compare the proposed method with state-of-the-art One-Shot NAS methods.

[The interplay between randomness and structure during learning in RNNs](#)

- Friedrich Schuessler, Francesca Mastrogiovanni, Alexis Dubreuil, Srdjan Ostojic, Omri Barak
- abstract@[open-review](#): Training recurrent neural networks (RNNs) on low-dimensional tasks has been widely used to model functional biological networks. However, the solutions found by learning and the effect of initial connectivity are not well understood. Here, we examine RNNs trained using gradient descent on different tasks inspired by the neuroscience literature. We find that the changes in recurrent connectivity can be described by low-rank matrices. This observation holds even in the presence of random initial connectivity, although this initial connectivity has full rank and significantly accelerates training. To understand the origin of these observations, we turn to an analytically tractable setting: training a linear RNN on a simpler task. We show how the low-dimensional task structure leads to low-rank changes to connectivity, and how random initial connectivity facilitates learning. Altogether, our study opens a new perspective to understand learning in RNNs in light of low-rank connectivity changes and the synergistic role of random initialization.

[A Generalized Neural Tangent Kernel Analysis for Two-layer Neural Networks](#)

- Zixiang Chen, Yuan Cao, Quanquan Gu, Tong Zhang
- abstract@[open-review](#): A recent breakthrough in deep learning theory shows that the training of over-parameterized deep neural networks can be characterized by a kernel function called \textit{neural tangent kernel} (NTK). However, it is known that this type of results does not perfectly match the practice, as NTK-based analysis requires the network weights to stay very close to their initialization throughout training, and cannot handle regularizers or gradient noises. In this paper, we provide a generalized neural tangent kernel analysis and show that noisy gradient descent with weight decay can still exhibit a ``kernel-like'' behavior. This implies that the training loss converges linearly up to a certain accuracy. We also establish a novel generalization error bound for two-layer neural networks trained by noisy gradient descent with weight decay.

[Instance-wise Feature Grouping](#)

- Aria Masoomi, Chieh Wu, Tingting Zhao, Zifeng Wang, Peter Castaldi, Jennifer Dy
- abstract@[open-review](#): In many learning problems, the domain scientist is often interested in discovering the groups of features that are redundant and are important for classification. Moreover, the features that belong to each group, and the important feature groups may vary per sample. But what do we mean by feature redundancy? In this paper, we formally define two types of redundancies using information theory: \textit{Representation} and \textit{Relevant redundancies}. We leverage these redundancies to design a formulation for instance-wise feature group discovery and reveal a theoretical guideline to help discover the appropriate number of groups. We approximate mutual information via a variational lower bound and learn the feature group and selector indicators with Gumbel-Softmax in optimizing our formulation. Experiments on synthetic data validate our theoretical claims. Experiments on MNIST, Fashion MNIST, and gene expression datasets show that our method discovers feature groups with high classification accuracies.

[Robust Disentanglement of a Few Factors at a Time using rPU-VAE](#)

- Benjamin Estermann, Markus Marks, Mehmet Fatih Yanik
- abstract@[open-review](#): Disentanglement is at the forefront of unsupervised learning, as disentangled representations of data improve generalization, interpretability, and performance in downstream tasks. Current unsupervised approaches remain inapplicable for real-world datasets since they are highly variable in their performance and fail to reach levels of disentanglement of (semi-)supervised approaches. We introduce population-based training (PBT) for improving consistency in training variational autoencoders (VAEs) and demonstrate the validity of this approach in a supervised setting (PBT-VAE). We then use Unsupervised Disentanglement Ranking (UDR) as an unsupervised heuristic to score models in our PBT-VAE training and show how models trained this way tend to consistently disentangle only a subset of the generative factors. Building on top of this observation we introduce the recursive rPU-VAE approach. We train the model until convergence, remove the learned factors from the dataset and reiterate. In doing so, we can label subsets of the dataset with the learned factors and consecutively use these labels to train one model that fully disentangles the whole dataset. With this approach, we show striking improvement in state-of-the-art unsupervised disentanglement performance and robustness across multiple datasets and metrics.

[PC-PG: Policy Cover Directed Exploration for Provable Policy Gradient Learning](#)

- Alekh Agarwal, Mikael Henaff, Sham Kakade, Wen Sun
- abstract@[open-review](#): Direct policy gradient methods for reinforcement learning are a successful approach for a variety of reasons: they are model free, they directly optimize the performance metric of interest, and they allow for richly parameterized policies. Their primary drawback is that, by being local in nature, they fail to adequately explore the environment. In contrast, while model-based approaches and Q-learning can, at least in theory, directly handle exploration through the use of optimism, their ability to handle model misspecification and function approximation is far less evident. This work introduces the the POLICY COVER GUIDED POLICY GRADIENT (PC- PG) algorithm, which provably balances the exploration vs. exploitation tradeoff using an ensemble of learned policies (the policy cover). PC-PG enjoys polynomial sample complexity and run time for both tabular MDPs and, more generally, linear MDPs in an infinite dimensional RKHS. Furthermore, PC-PG also has strong guarantees under model misspecification that go beyond the standard worst case L infinity assumptions; these include approximation guarantees for state aggregation under an average case error assumption, along with guarantees under a more general assumption where the approximation error under distribution shift is controlled. We complement the theory with empirical evaluation across a variety of domains in both reward-free and reward-driven settings.

[Group Contextual Encoding for 3D Point Clouds](#)

- Xu Liu, Chengtao Li, Jian Wang, Jingbo Wang, Boxin Shi, Xiaodong He
- abstract@[open-review](#): Global context is crucial for 3D point cloud scene understanding tasks. In this work, we extended the contextual encoding layer that was originally designed for 2D tasks to 3D Point Cloud scenarios. The encoding layer learns a set of code words in the feature space of the 3D point cloud to characterize the global semantic context, and then based on these code words, the method learns a global contextual descriptor to reweight the featuremaps accordingly. Moreover, compared to 2D scenarios, data sparsity becomes a major issue in 3D point cloud scenarios, and the performance of contextual encoding quickly saturates when the number of code words increases. To mitigate this problem, we further proposed a group contextual encoding method, which divides the channel into groups and then performs encoding on group-divided feature vectors. This method facilitates learning of global context in grouped subspace for 3D point clouds. We evaluate the effectiveness and generalizability of our method on three widely-studied 3D point cloud tasks. Experimental results have shown that the proposed method outperformed the VoteNet remarkably with 3 mAP on the benchmark of SUN-RGBD, with the metrics of mAP@ 0.25, and a much greater margin of 6.57 mAP on ScanNet with the metrics of mAP@ 0.5. Compared to the baseline of PointNet++, the proposed method leads to an accuracy of 86 %, outperforming the baseline by 1.5 %. Our proposed method have outperformed the non-grouping baseline methods across the board and establishes new state-of-the-art on these benchmarks.

[Latent Bandits Revisited](#)

- Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, Craig Boutilier
- abstract@[open-review](#): A latent bandit is a bandit problem where the learning agent knows reward distributions of arms conditioned on an unknown discrete latent state. The goal of the agent is to identify the latent state, after which it can act optimally. This setting is a natural midpoint between online and offline learning, where complex models can be learned offline and the agent identifies the latent state online. This is of high practical relevance, for instance in recommender systems. In this work, we propose general algorithms for latent bandits, based on both upper confidence bounds and Thompson sampling. The algorithms are contextual, and aware of model uncertainty and misspecification. We provide a unified theoretical analysis of our algorithms, which have lower regret than classic bandit policies when the number of latent states is smaller than actions. A comprehensive empirical study showcases the advantages of our approach.

[Is normalization indispensable for training deep neural network?](#)

- Jie Shao, Kai Hu, Changhu Wang, Xiangyang Xue, Bhiksha Raj
- abstract@[open-review](#): Normalization operations are widely used to train deep neural networks, and they can improve both convergence and generalization in most tasks. The theories for normalization's effectiveness and new forms of normalization have always been hot topics in research. To better understand normalization, one question can be whether normalization is indispensable for training deep neural network? In this paper, we study what would happen when normalization layers are removed from the network, and show how to train deep neural networks without normalization layers and without performance degradation. Our proposed method can achieve the same or even slightly better performance in a variety of tasks: image classification in ImageNet, object detection and segmentation in MS-COCO, video classification in Kinetics, and machine translation in WMT English-

German, etc. Our study may help better understand the role of normalization layers and can be a competitive alternative to normalization layers. Codes are available.

[Optimization and Generalization of Shallow Neural Networks with Quadratic Activation Functions](#)

- Stefano Sarao Mannelli, Eric Vanden-Eijnden, Lenka Zdeborová;
- abstract@[open-review](#): We study the dynamics of optimization and the generalization properties of one-hidden layer neural networks with quadratic activation function in the overparametrized regime where the layer width m is larger than the input dimension d . We consider a teacher-student scenario where the teacher has the same structure as the student with a hidden layer of smaller width $m \leq m$. We describe how the empirical loss landscape is affected by the number n of data samples and the width m of the teacher network. In particular we determine how the probability that there be no spurious minima on the empirical loss depends on n , d , and m^* , thereby establishing conditions under which the neural network can in principle recover the teacher. We also show that under the same conditions gradient descent dynamics on the empirical loss converges and leads to small generalization error, i.e. it enables recovery in practice. Finally we characterize the time-convergence rate of gradient descent in the limit of a large number of samples. These results are confirmed by numerical experiments.

[Intra Order-preserving Functions for Calibration of Multi-Class Neural Networks](#)

- Amir Rahimi, Amirreza Shaban, Ching-An Cheng, Richard Hartley, Byron Boots
- abstract@[open-review](#): Predicting calibrated confidence scores for multi-class deep networks is important for avoiding rare but costly mistakes. A common approach is to learn a post-hoc calibration function that transforms the output of the original network into calibrated confidence scores while maintaining the network's accuracy. However, previous post-hoc calibration techniques work only with simple calibration functions, potentially lacking sufficient representation to calibrate the complex function landscape of deep networks. In this work, we aim to learn general post-hoc calibration functions that can preserve the top-k predictions of any deep network. We call this family of functions intra order-preserving functions. We propose a new neural network architecture that represents a class of intra order-preserving functions by combining common neural network components. Additionally, we introduce order-invariant and diagonal sub-families, which can act as regularization for better generalization when the training data size is small. We show the effectiveness of the proposed method across a wide range of datasets and classifiers. Our method outperforms state-of-the-art post-hoc calibration methods, namely temperature scaling and Dirichlet calibration, in several evaluation metrics for the task.

[Linear Time Sinkhorn Divergences using Positive Features](#)

- Meyer Scetbon, Marco Cuturi
- abstract@[open-review](#): Although Sinkhorn divergences are now routinely used in data sciences to compare probability distributions, the computational effort required to compute them remains expensive, growing in general quadratically in the size n of the support of these distributions. Indeed, solving optimal transport (OT) with an entropic regularization requires computing a $n \times n$ kernel matrix (the neg-exponential of a $n \times n$ pairwise ground cost matrix) that is repeatedly applied to a vector. We propose to use instead ground costs of the form $c(x,y) = -\log(\text{dotp}(\varphi(x), \varphi(y)))$ where φ is a map from the ground space onto the positive orthant \mathbb{R}^n_+ , with $\|\varphi\|_1 \leq n$. This choice yields, equivalently, a kernel $k(x,y) = \text{dotp}(\varphi(x), \varphi(y))$, and ensures that the cost of Sinkhorn iterations scales as $O(nr)$. We show that usual cost functions can be approximated using this form. Additionally, we take advantage of the fact that our approach yields approximation that remain fully differentiable with respect to input distributions, as opposed to previously proposed adaptive low-rank approximations of the kernel matrix, to train a faster variant of OT-GAN~\cite{salimans2018improving}.

[VarGrad: A Low-Variance Gradient Estimator for Variational Inference](#)

- Lorenz Richter, Ayman Boustati, Nikolas Nüsken, Francisco Ruiz, Omer Deniz Akyildiz
- abstract@[open-review](#): We analyse the properties of an unbiased gradient estimator of the ELBO for variational inference, based on the score function method with leave-one-out control variates. We show that this gradient estimator can be obtained using a new loss, defined as the variance of the log-ratio between the exact posterior and the variational approximation, which we call the log-variance loss. Under certain conditions, the gradient of the log-variance loss equals the gradient of the (negative) ELBO. We show theoretically that this gradient estimator, which we call VarGrad due to its connection to the log-variance loss, exhibits lower variance than the score function method in certain settings, and that the leave-one-out control variate coefficients are close to the optimal ones. We empirically demonstrate that VarGrad offers a favourable variance versus computation trade-off compared to other state-of-the-art estimators on a discrete VAE.

[A Convolutional Auto-Encoder for Haplotype Assembly and Viral Quasispecies Reconstruction](#)

- Ziqi Ke, Haris Vikalo
- abstract@[open-review](#): Haplotype assembly and viral quasispecies reconstruction are challenging tasks concerned with analysis of genomic mixtures using sequencing data. High-throughput sequencing technologies generate enormous amounts of short fragments (reads) which essentially oversample components of a mixture; the representation redundancy enables reconstruction of the components (haplotypes, viral strains). The reconstruction problem, known to be NP-hard, boils down to grouping together reads originating from the same component in a mixture. Existing methods struggle to solve this problem with required level of accuracy and low runtimes; the problem is becoming increasingly more challenging as the number and length of the components increase. This paper proposes a read clustering method based on a convolutional auto-encoder designed to first project sequenced fragments to a low-dimensional space and then estimate the probability of the read origin using learned embedded features. The components are reconstructed by finding consensus sequences that agglomerate reads from the same origin. Mini-batch stochastic gradient descent and dimension reduction of reads allow the proposed method to efficiently deal with massive numbers of long reads. Experiments on simulated, semi-experimental and experimental data demonstrate the ability of the proposed method to accurately reconstruct haplotypes and viral quasispecies, often demonstrating superior performance compared to state-of-the-art methods. Source codes are available at <https://github.com/WuLoli/CAECseq>.

[Promoting Stochasticity for Expressive Policies via a Simple and Efficient Regularization Method](#)

- Qi Zhou, Yufei Kuang, Zherui Qiu, Houqiang Li, Jie Wang
- abstract@[open-review](#): Many recent reinforcement learning (RL) methods learn stochastic policies with entropy regularization for exploration and robustness. However, in continuous action spaces, integrating entropy regularization with expressive policies is challenging and usually requires complex inference procedures. To tackle this problem, we propose a novel regularization method that is compatible with a broad range of expressive policy architectures. An appealing feature is that, the estimation of our regularization terms is simple and efficient even when the policy distributions are unknown. We show that our approach can effectively promote the exploration in continuous action spaces. Based on our regularization, we propose an off-policy actor-critic algorithm. Experiments demonstrate that the proposed algorithm outperforms state-of-the-art regularized RL methods in continuous control tasks.

[Adversarial Counterfactual Learning and Evaluation for Recommender System](#)

- Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, Kannan Achan

- abstract@[open-review](#): The feedback data of recommender systems are often subject to what was exposed to the users; however, most learning and evaluation methods do not account for the underlying exposure mechanism. We first show in theory that applying supervised learning to detect user preferences may end up with inconsistent results in the absence of exposure information. The counterfactual propensity-weighting approach from causal inference can account for the exposure mechanism; nevertheless, the partial-observation nature of the feedback data can cause identifiability issues. We propose a principled solution by introducing a minimax empirical risk formulation. We show that the relaxation of the dual problem can be converted to an adversarial game between two recommendation models, where the opponent of the candidate model characterizes the underlying exposure mechanism. We provide learning bounds and conduct extensive simulation studies to illustrate and justify the proposed approach over a broad range of recommendation settings, which shed insights on the various benefits of the proposed approach.

[Memory-Efficient Learning of Stable Linear Dynamical Systems for Prediction and Control](#)

- Giorgos ('Yorgos') Mamakoukas, Orest Xherija, Todd Murphrey
- abstract@[open-review](#): Learning a stable Linear Dynamical System (LDS) from data involves creating models that both minimize reconstruction error and enforce stability of the learned representation. We propose a novel algorithm for learning stable LDSs. Using a recent characterization of stable matrices, we present an optimization method that ensures stability at every step and iteratively improves the reconstruction error using gradient directions derived in this paper. When applied to LDSs with inputs, our approach---in contrast to current methods for learning stable LDSs---updates both the state and control matrices, expanding the solution space and allowing for models with lower reconstruction error. We apply our algorithm in simulations and experiments to a variety of problems, including learning dynamic textures from image sequences and controlling a robotic manipulator. Compared to existing approaches, our proposed method achieves an $\text{orders-of-magnitude}$ improvement in reconstruction error and superior results in terms of control performance. In addition, it is provably more memory efficient, with an $\mathcal{O}(n^2)$ space complexity compared to $\mathcal{O}(n^4)$ of competing alternatives, thus scaling to higher-dimensional systems when the other methods fail. The code of the proposed algorithm and animations of the results can be found at <https://github.com/giorgosmamakoukas/MemoryEfficientStableLDS>.

[Evolving Normalization-Activation Layers](#)

- Hanxiao Liu, Andy Brock, Karen Simonyan, Quoc Le
- abstract@[open-review](#): Normalization layers and activation functions are fundamental components in deep networks and typically co-locate with each other. Here we propose to design them using an automated approach. Instead of designing them separately, we unify them into a single tensor-to-tensor computation graph, and evolve its structure starting from basic mathematical functions. Examples of such mathematical functions are addition, multiplication and statistical moments. The use of low-level mathematical functions, in contrast to the use of high-level modules in mainstream NAS, leads to a highly sparse and large search space which can be challenging for search methods. To address the challenge, we develop efficient rejection protocols to quickly filter out candidate layers that do not work well. We also use multi-objective evolution to optimize each layer's performance across many architectures to prevent overfitting. Our method leads to the discovery of EvoNorms, a set of new normalization-activation layers with novel, and sometimes surprising structures that go beyond existing design patterns. For example, some EvoNorms do not assume that normalization and activation functions must be applied sequentially, nor need to center the feature maps, nor require explicit activation functions. Our experiments show that EvoNorms work well on image classification models including ResNets, MobileNets and EfficientNets but also transfer well to Mask R-CNN with FPN/SpineNet for instance segmentation and to BigGAN for image synthesis, outperforming BatchNorm and GroupNorm based layers in many cases.

[ScaleCom: Scalable Sparsified Gradient Compression for Communication-Efficient Distributed Training](#)

- Chia-Yu Chen, Jiamin Ni, Songtao Lu, Xiaodong Cui, Pin-Yu Chen, Xiao Sun, Naigang Wang, Swagath Venkataramani, Vijayalakshmi (Viji) Srinivasan, Wei Zhang, Kailash Gopalakrishnan
- abstract@[open-review](#): Large-scale distributed training of Deep Neural Networks (DNNs) on state-of-the-art platforms are expected to be severely communication constrained. To overcome this limitation, numerous gradient compression techniques have been proposed and have demonstrated high compression ratios. However, most existing compression methods do not scale well to large scale distributed systems (due to gradient build-up) and / or lack evaluations in large datasets. To mitigate these issues, we propose a new compression technique, Scalable Sparsified Gradient Compression (ScaleComp), that (i) leverages similarity in the gradient distribution amongst learners to provide a commutative compressor and keep communication cost constant to worker number and (ii) includes low-pass filter in local gradient accumulations to mitigate the impacts of large batch size training and significantly improve scalability. Using theoretical analysis, we show that ScaleComp provides favorable convergence guarantees and is compatible with gradient all-reduce techniques. Furthermore, we experimentally demonstrate that ScaleComp has small overheads, directly reduces gradient traffic and provides high compression rates (70-150X) and excellent scalability (up to 64-80 learners and 10X larger batch sizes over normal training) across a wide range of applications (image, language, and speech) without significant accuracy loss.

[RelationNet++: Bridging Visual Representations for Object Detection via Transformer Decoder](#)

- Cheng Chi, Fangyun Wei, Han Hu
- abstract@[open-review](#): Existing object detection frameworks are usually built on a single format of object/part representation, i.e., anchor/proposal rectangle boxes in RetinaNet and Faster R-CNN, center points in FCOS and RepPoints, and corner points in CornerNet. While these different representations usually drive the frameworks to perform well in different aspects, e.g., better classification or finer localization, it is in general difficult to combine these representations in a single framework to make good use of each strength, due to the heterogeneous or non-grid feature extraction by different representations. This paper presents an attention-based decoder module similar as that in Transformer~\cite{vaswani2017attention} to bridge other representations into a typical object detector built on a single representation format, in an end-to-end fashion. The other representations act as a set of \emph{key} instances to strengthen the main \emph{query} representation features in the vanilla detectors. Novel techniques are proposed towards efficient computation of the decoder module, including a \emph{key sampling} approach and a \emph{shared location embedding} approach. The proposed module is named \emph{bridging visual representations} (BVR). It can perform in-place and we demonstrate its broad effectiveness in bridging other representations into prevalent object detection frameworks, including RetinaNet, Faster R-CNN, FCOS and ATSS, where about \$1.5\sim 3.0\$ AP improvements are achieved. In particular, we improve a state-of-the-art framework with a strong backbone by about \$2.0\$ AP, reaching \$52.7\$ AP on COCO test-dev. The resulting network is named RelationNet++. The code is available at \url{https://github.com/microsoft/RelationNet2}.

[Efficient Learning of Discrete Graphical Models](#)

- Marc Vuffray, Sidhant Misra, Andrey Lokhov
- abstract@[open-review](#): Graphical models are useful tools for describing structured high-dimensional probability distributions. Development of efficient algorithms for learning graphical models with least amount of data remains an active research topic. Reconstruction of graphical models that describe the statistics of discrete variables is a particularly challenging problem, for which the maximum likelihood approach is intractable. In this work, we provide the first sample-efficient method based on the Interaction Screening framework that allows one to provably learn fully general discrete factor models with node-specific discrete alphabets and multi-body interactions, specified in an arbitrary basis. We identify a single condition related to model parametrization that leads to rigorous guarantees on the recovery of model structure and parameters in any error norm, and is readily verifiable for a large class of models. Importantly, our bounds make explicit distinction between parameters that are proper to the model and priors used as an input to the algorithm. Finally, we show that the Interaction Screening framework includes all models previously considered in the literature as special cases, and for which our analysis shows a systematic improvement in sample complexity.

[Near-Optimal SQ Lower Bounds for Agnostically Learning Halfspaces and ReLUs under Gaussian Marginals](#)

- Ilias Diakonikolas, Daniel Kane, Nikos Zarifis
- abstract@[open-review](#): We study the fundamental problems of agnostically learning halfspaces and ReLUs under Gaussian marginals. In the former problem, given labeled examples (\mathbf{b}_x, y) from an unknown distribution on $\mathbb{R}^d \times \{\pm 1\}$, whose marginal distribution on \mathbf{b}_x is the standard Gaussian and the labels y can be arbitrary, the goal is to output a hypothesis with 0-1 loss $\text{opt} + \epsilon$, where opt is the 0-1 loss of the best-fitting halfspace. In the latter problem, given labeled examples (\mathbf{b}_x, y) from an unknown distribution on $\mathbb{R}^d \times \mathbb{R}$, whose marginal distribution on \mathbf{b}_x is the standard Gaussian and the labels y can be arbitrary, the goal is to output a hypothesis with square loss $\text{opt} + \epsilon$, where opt is the square loss of the best-fitting ReLU. We prove Statistical Query (SQ) lower bounds of $d^{\{\text{poly}(1/\epsilon)\}}$ for both of these problems. Our SQ lower bounds provide strong evidence that current upper bounds for these tasks are essentially best possible.

[Neurosymbolic Transformers for Multi-Agent Communication](#)

- Jeevana Priya Inala, Yichen Yang, James Paulos, Yewen Pu, Osbert Bastani, Vijay Kumar, Martin Rinard, Armando Solar-Lezama
- abstract@[open-review](#): We study the problem of inferring communication structures that can solve cooperative multi-agent planning problems while minimizing the amount of communication. We quantify the amount of communication as the maximum degree of the communication graph; this metric captures settings where agents have limited bandwidth. Minimizing communication is challenging due to the combinatorial nature of both the decision space and the objective; for instance, we cannot solve this problem by training neural networks using gradient descent. We propose a novel algorithm that synthesizes a control policy that combines a programmatic communication policy used to generate the communication graph with a transformer policy network used to choose actions. Our algorithm first trains the transformer policy, which implicitly generates a "soft" communication graph; then, it synthesizes a programmatic communication policy that "hardens" this graph, forming a neurosymbolic transformer. Our experiments demonstrate how our approach can synthesize policies that generate low-degree communication graphs while maintaining near-optimal performance.

[Fairness in Streaming Submodular Maximization: Algorithms and Hardness](#)

- Marwa El Halabi, Slobodan Mitrović, Ashkan Norouzi-Fard, Jakab Tardos, Jakub M. Tarnawski
- abstract@[open-review](#): Submodular maximization has become established as the method of choice for the task of selecting representative and diverse summaries of data. However, if datapoints have sensitive attributes such as gender or age, such machine learning algorithms, left unchecked, are known to exhibit bias: under- or over-representation of particular groups. This has made the design of fair machine learning algorithms increasingly important. In this work we address the question: Is it possible to create fair summaries for massive datasets? To this end, we develop the first streaming approximation algorithms for submodular maximization under fairness constraints, for both monotone and non-monotone functions. We validate our findings empirically on exemplar-based clustering, movie recommendation, DPP-based summarization, and maximum coverage in social networks, showing that fairness constraints do not significantly impact utility.

[Smoothed Geometry for Robust Attribution](#)

- Zifan Wang, Haofan Wang, Shakul Ramkumar, Piotr Mardziel, Matt Fredrikson, Anupam Datta
- abstract@[open-review](#): Feature attributions are a popular tool for explaining the behavior of Deep Neural Networks (DNNs), but have recently been shown to be vulnerable to attacks that produce divergent explanations for nearby inputs. This lack of robustness is especially problematic in high-stakes applications where adversarially-manipulated explanations could impair safety and trustworthiness. Building on a geometric understanding of these attacks presented in recent work, we identify Lipschitz continuity conditions on models' gradient that lead to robust gradient-based attributions, and observe that smoothness may also be related to the ability of an attack to transfer across multiple attribution methods. To mitigate these attacks in practice, we propose an inexpensive regularization method that promotes these conditions in DNNs, as well as a stochastic smoothing technique that does not require re-training. Our experiments on a range of image models demonstrate that both of these mitigations consistently improve attribution robustness, and confirm the role that smooth geometry plays in these attacks on real, large-scale models.

[Fast Adversarial Robustness Certification of Nearest Prototype Classifiers for Arbitrary Seminorms](#)

- Sascha Saralajew, Lars Holdijk, Thomas Villmann
- abstract@[open-review](#): Methods for adversarial robustness certification aim to provide an upper bound on the test error of a classifier under adversarial manipulation of its input. Current certification methods are computationally expensive and limited to attacks that optimize the manipulation with respect to a norm. We overcome these limitations by investigating the robustness properties of Nearest Prototype Classifiers (NPCs) like learning vector quantization and large margin nearest neighbor. For this purpose, we study the hypothesis margin. We prove that if NPCs use a dissimilarity measure induced by a seminorm, the hypothesis margin is a tight lower bound on the size of adversarial attacks and can be calculated in constant time—this provides the first adversarial robustness certificate calculable in reasonable time. Finally, we show that each NPC trained by a triplet loss maximizes the hypothesis margin and is therefore optimized for adversarial robustness. In the presented evaluation, we demonstrate that NPCs optimized for adversarial robustness are competitive with state-of-the-art methods and set a new benchmark with respect to computational complexity for robustness certification.

[Multi-agent active perception with prediction rewards](#)

- Mikko Lauri, Frans Oliehoek
- abstract@[open-review](#): Multi-agent active perception is a task where a team of agents cooperatively gathers observations to compute a joint estimate of a hidden variable. The task is decentralized and the joint estimate can only be computed after the task ends by fusing observations of all agents. The objective is to maximize the accuracy of the estimate. The accuracy is quantified by a centralized prediction reward determined by a centralized decision-maker who perceives the observations gathered by all agents after the task ends. In this paper, we model multi-agent active perception as a decentralized partially observable Markov decision process (Dec-POMDP) with a convex centralized prediction reward. We prove that by introducing individual prediction actions for each agent, the problem is converted into a standard Dec-POMDP with a decentralized prediction reward. The loss due to decentralization is bounded, and we give a sufficient condition for when it is zero. Our results allow application of any Dec-POMDP solution algorithm to multi-agent active perception problems, and enable planning to reduce uncertainty without explicit computation of joint estimates. We demonstrate the empirical usefulness of our results by applying a standard Dec-POMDP algorithm to multi-agent active perception problems, showing increased scalability in the planning horizon.

[A Local Temporal Difference Code for Distributional Reinforcement Learning](#)

- Pablo Tano, Peter Dayan, Alexandre Pouget
- abstract@[open-review](#): Recent theoretical and experimental results suggest that the dopamine system implements distributional temporal difference backups, allowing learning of the entire distributions of the long-run values of states rather than just their expected values. However, the distributional codes explored so far rely on a complex imputation step which crucially relies on spatial non-locality: in order to compute reward prediction errors, units must know not only their own state but also the states of the other units. It is far from clear how these steps could be implemented in realistic neural circuits. Here, we introduce the Laplace code: a local temporal difference code for distributional reinforcement learning that is representationally powerful and computationally straightforward. The code decomposes value distributions and prediction errors across three separated dimensions: reward magnitude

(related to distributional quantiles), temporal discounting (related to the Laplace transform of future rewards) and time horizon (related to eligibility traces). Besides lending itself to a local learning rule, the decomposition recovers the temporal evolution of the immediate reward distribution, indicating all possible rewards at all future times. This increases representational capacity and allows for temporally-flexible computations that immediately adjust to changing horizons or discount factors.

[Learning with Optimized Random Features: Exponential Speedup by Quantum Machine Learning without Sparsity and Low-Rank Assumptions](#)

- Hayata Yamasaki, Sathyawageeswar Subramanian, Sho Sonoda, Masato Koashi
- abstract@[open-review](#): Kernel methods augmented with random features give scalable algorithms for learning from big data. But it has been computationally hard to sample random features according to a probability distribution that is optimized for the data, so as to minimize the required number of features for achieving the learning to a desired accuracy. Here, we develop a quantum algorithm for sampling from this optimized distribution over features, in runtime $O(D)$ that is linear in the dimension D of the input data. Our algorithm achieves an exponential speedup in D compared to any known classical algorithm for this sampling task. In contrast to existing quantum machine learning algorithms, our algorithm circumvents sparsity and low-rank assumptions and thus has wide applicability. We also show that the sampled features can be combined with regression by stochastic gradient descent to achieve the learning without canceling out our exponential speedup. Our algorithm based on sampling optimized random features leads to an accelerated framework for machine learning that takes advantage of quantum computers.

[CaSPR: Learning Canonical Spatiotemporal Point Cloud Representations](#)

- Davis Rempe, Tolga Birdal, Yongheng Zhao, Zan Gojcic, Srinath Sridhar, Leonidas J. Guibas
- abstract@[open-review](#): We propose CaSPR, a method to learn object-centric Canonical Spatiotemporal Point Cloud Representations of dynamically moving or evolving objects. Our goal is to enable information aggregation over time and the interrogation of object state at any spatiotemporal neighborhood in the past, observed or not. Different from previous work, CaSPR learns representations that support spacetime continuity, are robust to variable and irregularly spacetime-sampled point clouds, and generalize to unseen object instances. Our approach divides the problem into two subtasks. First, we explicitly encode time by mapping an input point cloud sequence to a spatiotemporally-canonicalized object space. We then leverage this canonicalization to learn a spatiotemporal latent representation using neural ordinary differential equations and a generative model of dynamically evolving shapes using continuous normalizing flows. We demonstrate the effectiveness of our method on several applications including shape reconstruction, camera pose estimation, continuous spatiotemporal sequence reconstruction, and correspondence estimation from irregularly or intermittently sampled observations.

[Deep Automodulators](#)

- Ari Heljakka, Yuxin Hou, Juho Kannala, Arno Solin
- abstract@[open-review](#): We introduce a new category of generative autoencoders called automodulators. These networks can faithfully reproduce individual real-world input images like regular autoencoders, but also generate a fused sample from an arbitrary combination of several such images, allowing instantaneous "style-mixing" and other new applications. An automodulator decouples the data flow of decoder operations from statistical properties thereof and uses the latent vector to modulate the former by the latter, with a principled approach for mutual disentanglement of decoder layers. Prior work has explored similar decoder architecture with GANs, but their focus has been on random sampling. A corresponding autoencoder could operate on real input images. For the first time, we show how to train such a general-purpose model with sharp outputs in high resolution, using novel training techniques, demonstrated on four image data sets. Besides style-mixing, we show state-of-the-art results in autoencoder comparison, and visual image quality nearly indistinguishable from state-of-the-art GANs. We expect the automodulator variants to become a useful building block for image applications and other data domains.

[Convolutional Tensor-Train LSTM for Spatio-Temporal Learning](#)

- Jiahao Su, Wonmin Byeon, Jean Kossaifi, Furong Huang, Jan Kautz, Anima Anandkumar
- abstract@[open-review](#): Learning from spatio-temporal data has numerous applications such as human-behavior analysis, object tracking, video compression, and physics simulation. However, existing methods still perform poorly on challenging video tasks such as long-term forecasting. This is because these kinds of challenging tasks require learning long-term spatio-temporal correlations in the video sequence. In this paper, we propose a higher-order convolutional LSTM model that can efficiently learn these correlations, along with a succinct representations of the history. This is accomplished through a novel tensor train module that performs prediction by combining convolutional features across time. To make this feasible in terms of computation and memory requirements, we propose a novel convolutional tensor-train decomposition of the higher-order model. This decomposition reduces the model complexity by jointly approximating a sequence of convolutional kernels as a low-rank tensor-train factorization. As a result, our model outperforms existing approaches, but uses only a fraction of parameters, including the baseline models. Our results achieve state-of-the-art performance in a wide range of applications and datasets, including the multi-steps video prediction on the Moving-MNIST-2 and KTH action datasets as well as early activity recognition on the Something-Something V2 dataset.

[The Potts-Ising model for discrete multivariate data](#)

- Zahra Razaee, Arash Amini
- abstract@[open-review](#): Modeling dependencies in multivariate discrete data is a challenging problem, especially in high dimensions. The Potts model is a versatile such model, suitable when each coordinate is a categorical variable. However, the full Potts model has too many parameters to be accurately fit when the number of categories is large. We introduce a variation on the Potts model that allows for general categorical marginals and Ising-type multivariate dependence. This reduces the number of parameters from $\Omega(d^2 K^2)$ in the full Potts model to $O(d^2 + Kd)$, where K is the number of categories and d is the dimension of the data. We show that the complexity of fitting this new Potts-Ising model is the same as that of an Ising model. In particular, adopting the neighborhood regression framework, the model can be fit by solving d separate logistic regressions. We demonstrate the ability of the model to capture multivariate dependencies by comparing with existing approaches.

[Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech](#)

- Shailee Jain, Vy Vo, Shivangi Mahto, Amanda LeBel, Javier S. Turek, Alexander Huth
- abstract@[open-review](#): Natural language contains information at multiple timescales. To understand how the human brain represents this information, one approach is to build encoding models that predict fMRI responses to natural language using representations extracted from neural network language models (LMs). However, these LM-derived representations do not explicitly separate information at different timescales, making it difficult to interpret the encoding models. In this work we construct interpretable multi-timescale representations by forcing individual units in an LSTM LM to integrate information over specific temporal scales. This allows us to explicitly and directly map the timescale of information encoded by each individual fMRI voxel. Further, the standard fMRI encoding procedure does not account for varying temporal properties in the encoding features. We modify the procedure so that it can capture both short- and long-timescale information. This approach outperforms other encoding models, particularly for voxels that represent long-timescale information. It also provides a finer-grained map of timescale information in the human language pathway. This serves as a framework for future work investigating temporal hierarchies across artificial and biological language systems.

Group-Fair Online Allocation in Continuous Time

- Semih Cayci, Swati Gupta, Atilla Eryilmaz
- abstract@[open-review](#): The theory of discrete-time online learning has been successfully applied in many problems that involve sequential decision-making under uncertainty. However, in many applications including contractual hiring in online freelancing platforms and server allocation in cloud computing systems, the outcome of each action is observed only after a random and action-dependent time. Furthermore, as a consequence of certain ethical and economic concerns, the controller may impose deadlines on the completion of each task, and require fairness across different groups in the allocation of total time budget $\$B\$$. In order to address these applications, we consider continuous-time online learning problem with fairness considerations, and present a novel framework based on continuous-time utility maximization. We show that this formulation recovers reward-maximizing, max-min fair and proportionally fair allocation rules across different groups as special cases. We characterize the optimal offline policy, which allocates the total time between different actions in an optimally fair way (as defined by the utility function), and impose deadlines to maximize time-efficiency. In the absence of any statistical knowledge, we propose a novel online learning algorithm based on dual ascent optimization for time averages, and prove that it achieves $\tilde{O}(B^{-1/2})$ regret bound.

Decentralized TD Tracking with Linear Function Approximation and its Finite-Time Analysis

- Gang Wang, Songtao Lu, Georgios Giannakis, Gerald Tesauro, Jian Sun
- abstract@[open-review](#): The present contribution deals with decentralized policy evaluation in multi-agent Markov decision processes using temporal-difference (TD) methods with linear function approximation for scalability. The agents cooperate to estimate the value function of such a process by observing continual state transitions of a shared environment over the graph of interconnected nodes (agents), along with locally private rewards. Different from existing consensus-type TD algorithms, the approach here develops a simple decentralized TD tracker by wedging TD learning with gradient tracking techniques. The non-asymptotic properties of the novel TD tracker are established for both independent and identically distributed (i.i.d.) as well as Markovian transitions through a unifying multistep Lyapunov analysis. In contrast to the prior art, the novel algorithm forgoes the limiting error bounds on the number of agents, which endows it with performance comparable to that of centralized TD methods that are the sharpest known to date.

Understanding Gradient Clipping in Private SGD: A Geometric Perspective

- Xiangyi Chen, Steven Z. Wu, Mingyi Hong
- abstract@[open-review](#): Deep learning models are increasingly popular in many machine learning applications where the training data may contain sensitive information. To provide formal and rigorous privacy guarantee, many learning systems now incorporate differential privacy by training their models with (differentially) private SGD. A key step in each private SGD update is gradient clipping that shrinks the gradient of an individual example whenever its L_2 norm exceeds a certain threshold. We first demonstrate how gradient clipping can prevent SGD from converging to a stationary point. We then provide a theoretical analysis on private SGD with gradient clipping. Our analysis fully characterizes the clipping bias on the gradient norm, which can be upper bounded by the Wasserstein distance between the gradient distribution and a geometrically symmetric distribution. Our empirical evaluation further suggests that the gradient distributions along the trajectory of private SGD indeed exhibit such symmetric structure. Together, our results provide an explanation why private SGD with gradient clipping remains effective in practice despite its potential clipping bias. Finally, we develop a new perturbation-based technique that can provably correct the clipping bias even for instances with highly asymmetric gradient distributions.

O(n) Connections are Expressive Enough: Universal Approximability of Sparse Transformers

- Chulhee Yun, Yin-Wen Chang, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, Sanjiv Kumar
- abstract@[open-review](#): Recently, Transformer networks have redefined the state of the art in many NLP tasks. However, these models suffer from quadratic computational cost in the input sequence length n to compute pairwise attention in each layer. This has prompted recent research into sparse Transformers that sparsify the connections in the attention layers. While empirically promising for long sequences, fundamental questions remain unanswered: Can sparse Transformers approximate any arbitrary sequence-to-sequence function, similar to their dense counterparts? How does the sparsity pattern and the sparsity level affect their performance? In this paper, we address these questions and provide a unifying framework that captures existing sparse attention models. We propose sufficient conditions under which we prove that a sparse attention model can universally approximate any sequence-to-sequence function. Surprisingly, our results show that sparse Transformers with only $O(n)$ connections per attention layer can approximate the same function class as the dense model with n^2 connections. Lastly, we present experiments comparing different patterns/levels of sparsity on standard NLP tasks.

Identifying signal and noise structure in neural population activity with Gaussian process factor models

- Stephen Keeley, Mikio Aoi, Yiyi Yu, Spencer Smith, Jonathan W. Pillow
- abstract@[open-review](#): Neural datasets often contain measurements of neural activity across multiple trials of a repeated stimulus or behavior. An important problem in the analysis of such datasets is to characterize systematic aspects of neural activity that carry information about the repeated stimulus or behavior of interest, which can be considered "signal", and to separate them from the trial-to-trial fluctuations in activity that are not time-locked to the stimulus, which for purposes of such analyses can be considered "noise". Gaussian Process Factor models provide a powerful tool for identifying shared structure in high-dimensional neural data. However, they have not yet been adapted to the problem of characterizing signal and noise in multi-trial datasets. Here we address this shortcoming by proposing ``signal-noise'' Poisson-spiking Gaussian Process Factor Analysis (SNP-GPFA), a flexible latent variable model that resolves signal and noise latent structure in neural population spiking activity. To learn the parameters of our model, we introduce a Fourier-domain black box variational inference method that quickly identifies smooth latent structure. The resulting model reliably uncovers latent signal and trial-to-trial noise-related fluctuations in large-scale recordings. We use this model to show that in monkey V1, noise fluctuations perturb neural activity within a subspace orthogonal to signal activity, suggesting that trial-by-trial noise does not interfere with signal representations. Finally, we extend the model to capture statistical dependencies across brain regions in multi-region data. We show that in mouse visual cortex, models with shared noise across brain regions out-perform models with independent per-region noise.

Equivariant Networks for Hierarchical Structures

- Renhao Wang, Marjan Albooyeh, Siamak Ravanbakhsh
- abstract@[open-review](#): While using invariant and equivariant maps, it is possible to apply deep learning to a range of primitive data structures, a formalism for dealing with hierarchy is lacking. This is a significant issue because many practical structures are hierarchies of simple building blocks; some examples include sequences of sets, graphs of graphs, or multiresolution images. Observing that the symmetry of a hierarchical structure is the ``wreath product'' of symmetries of the building blocks, we express the equivariant map for the hierarchy using an intuitive combination of the equivariant linear layers of the building blocks. More generally, we show that any equivariant map for the hierarchy has this form. To demonstrate the effectiveness of this approach to model design, we consider its application in the semantic segmentation of point-cloud data. By voxelizing the point cloud, we impose a hierarchy of translation and permutation symmetries on the data and report state-of-the-art on {semantic3d}, {s3dis}, and {vkitti}, that include some of the largest real-world point-cloud benchmarks.

MinMax Methods for Optimal Transport and Beyond: Regularization, Approximation and Numerics

- Luca De Gennaro Aquino, Stephan Eckstein
- abstract@[open-review](#): We study MinMax solution methods for a general class of optimization problems related to (and including) optimal transport. Theoretically, the focus is on fitting a large class of problems into a single MinMax framework and generalizing regularization techniques known from classical optimal transport. We show that regularization techniques justify the utilization of neural networks to solve such problems by proving approximation theorems and illustrating fundamental issues if no regularization is used. We further study the relation to the literature on generative adversarial nets, and analyze which algorithmic techniques used therein are particularly suitable to the class of problems studied in this paper. Several numerical experiments showcase the generality of the setting and highlight which theoretical insights are most beneficial in practice.

[A Discrete Variational Recurrent Topic Model without the Reparametrization Trick](#)

- Mehdi Rezaee, Francis Ferraro
- abstract@[open-review](#): We show how to learn a neural topic model with discrete random variables---one that explicitly models each word's assigned topic---using neural variational inference that does not rely on stochastic backpropagation to handle the discrete variables. The model we utilize combines the expressive power of neural methods for representing sequences of text with the topic model's ability to capture global, thematic coherence. Using neural variational inference, we show improved perplexity and document understanding across multiple corpora. We examine the effect of prior parameters both on the model and variational parameters, and demonstrate how our approach can compete and surpass a popular topic model implementation on an automatic measure of topic quality.

[Transferable Graph Optimizers for ML Compilers](#)

- Yanqi Zhou, Sudip Roy, Amirali Abdolrashidi, Daniel Wong, Peter Ma, Qiumin Xu, Hanxiao Liu, Phitchaya Phothilimtha, Shen Wang, Anna Goldie, Azalia Mirhoseini, James Laudon
- abstract@[open-review](#): Most compilers for machine learning (ML) frameworks need to solve many correlated optimization problems to generate efficient machine code. Current ML compilers rely on heuristics based algorithms to solve these optimization problems one at a time. However, this approach is not only hard to maintain but often leads to sub-optimal solutions especially for newer model architectures. Existing learning based approaches in the literature are sample inefficient, tackle a single optimization problem, and do not generalize to unseen graphs making them infeasible to be deployed in practice. To address these limitations, we propose an end-to-end, transferable deep reinforcement learning method for computational graph optimization (GO), based on a scalable sequential attention mechanism over an inductive graph neural network. GO generates decisions on the entire graph rather than on each individual node autoregressively, drastically speeding up the search compared to prior methods. Moreover, we propose recurrent attention layers to jointly optimize dependent graph optimization tasks and demonstrate 33%-60% speedup on three graph optimization tasks compared to TensorFlow default optimization. On a diverse set of representative graphs consisting of up to 80,000 nodes, including Inception-v3, Transformer-XL, and WaveNet, GO achieves on average 21% improvement over human experts and 18% improvement over the prior state of the art with 15x faster convergence, on a device placement task evaluated in real systems.

[Learning with Operator-valued Kernels in Reproducing Kernel Krein Spaces](#)

- Akash Saha, Balamurugan Palaniappan
- abstract@[open-review](#): Operator-valued kernels have shown promise in supervised learning problems with functional inputs and functional outputs. The crucial (and possibly restrictive) assumption of positive definiteness of operator-valued kernels has been instrumental in developing efficient algorithms. In this work, we consider operator-valued kernels which might not be necessarily positive definite. To tackle the indefiniteness of operator-valued kernels, we harness the machinery of Reproducing Kernel Krein Spaces (RKKS) of function-valued functions. A representer theorem is illustrated which yields a suitable loss stabilization problem for supervised learning with function-valued inputs and outputs. Analysis of generalization properties of the proposed framework is given. An iterative Operator based Minimum Residual (OpMINRES) algorithm is proposed for solving the loss stabilization problem. Experiments with indefinite operator-valued kernels on synthetic and real data sets demonstrate the utility of the proposed approach.

[Learning Bounds for Risk-sensitive Learning](#)

- Jaeho Lee, Sejun Park, Jinwoo Shin
- abstract@[open-review](#): In risk-sensitive learning, one aims to find a hypothesis that minimizes a risk-averse (or risk-seeking) measure of loss, instead of the standard expected loss. In this paper, we propose to study the generalization properties of risk-sensitive learning schemes whose optimand is described via optimized certainty equivalents (OCE): our general scheme can handle various known risks, e.g., the entropic risk, mean-variance, and conditional value-at-risk, as special cases. We provide two learning bounds on the performance of empirical OCE minimizer. The first result gives an OCE guarantee based on the Rademacher average of the hypothesis space, which generalizes and improves existing results on the expected loss and the conditional value-at-risk. The second result, based on a novel variance-based characterization of OCE, gives an expected loss guarantee with a suppressed dependence on the smoothness of the selected OCE. Finally, we demonstrate the practical implications of the proposed bounds via exploratory experiments on neural networks.

[Simplifying Hamiltonian and Lagrangian Neural Networks via Explicit Constraints](#)

- Marc Finzi, Ke Alexander Wang, Andrew G. Wilson
- abstract@[open-review](#): Reasoning about the physical world requires models that are endowed with the right inductive biases to learn the underlying dynamics. Recent works improve generalization for predicting trajectories by learning the Hamiltonian or Lagrangian of a system rather than the differential equations directly. While these methods encode the constraints of the systems using generalized coordinates, we show that embedding the system into Cartesian coordinates and enforcing the constraints explicitly with Lagrange multipliers dramatically simplifies the learning problem. We introduce a series of challenging chaotic and extended-body systems, including systems with \$N\$-pendulums, spring coupling, magnetic fields, rigid rotors, and gyroscopes, to push the limits of current approaches. Our experiments show that Cartesian coordinates with explicit constraints lead to a 100x improvement in accuracy and data efficiency.

[Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency](#)

- Robert Geirhos, Kristof Meding, Felix A. Wichmann
- abstract@[open-review](#): A central problem in cognitive science and behavioural neuroscience as well as in machine learning and artificial intelligence research is to ascertain whether two or more decision makers---be they brains or algorithms---use the same strategy. Accuracy alone cannot distinguish between strategies: two systems may achieve similar accuracy with very different strategies. The need to differentiate beyond accuracy is particularly pressing if two systems are at or near ceiling performance, like Convolutional Neural Networks (CNNs) and humans on visual object recognition. Here we introduce trial-by-trial error consistency, a quantitative analysis for measuring whether two decision making systems systematically make errors on the same inputs. Making consistent errors on a trial-by-trial basis is a necessary condition if we want to ascertain similar processing strategies between decision makers. Our analysis is applicable to compare algorithms with algorithms, humans with humans, and algorithms with humans. When applying error consistency to visual object recognition we obtain three main findings: (1.) Irrespective of architecture, CNNs are remarkably consistent with one another. (2.) The consistency between CNNs and human observers, however, is little above what can be expected by chance alone---indicating that humans and CNNs are likely implementing very different strategies. (3.) CORnet-S, a recurrent model termed the "current best model of the primate ventral visual stream", fails to capture essential characteristics of human behavioural data and behaves essentially like a standard purely feedforward

ResNet-50 in our analysis; highlighting that certain behavioural failure cases are not limited to feedforward models. Taken together, error consistency analysis suggests that the strategies used by human and machine vision are still very different---but we envision our general-purpose error consistency analysis to serve as a fruitful tool for quantifying future progress.

[Provably Efficient Reinforcement Learning with Kernel and Neural Function Approximations](#)

- Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, Michael Jordan
- abstract@[open-review](#): Reinforcement learning (RL) algorithms combined with modern function approximators such as kernel functions and deep neural networks have achieved significant empirical successes in large-scale application problems with a massive number of states. From a theoretical perspective, however, RL with functional approximation poses a fundamental challenge to developing algorithms with provable computational and statistical efficiency, due to the need to take into consideration both the exploration-exploitation tradeoff that is inherent in RL and the bias-variance tradeoff that is innate in statistical estimation. To address such a challenge, focusing on the episodic setting where the action-value functions are represented by a kernel function or over-parametrized neural network, we propose the first provable RL algorithm with both polynomial runtime and sample complexity, without additional assumptions on the data-generating model. In particular, for both the kernel and neural settings, we prove that an optimistic modification of the least-squares value iteration algorithm incurs an $\tilde{O}(\delta_{\mathcal{C}} H^2 \sqrt{T})$ regret, where $\delta_{\mathcal{C}}$ characterizes the intrinsic complexity of the function class \mathcal{C} , H is the length of each episode, and T is the total number of episodes. Our regret bounds are independent of the number of states and therefore even allows it to diverge, which exhibits the benefit of function approximation.

[Constant-Expansion Suffices for Compressed Sensing with Generative Priors](#)

- Constantinos Daskalakis, Dhruv Rohatgi, Emmanouil Zampetakis
- abstract@[open-review](#): Generative neural networks have been empirically found very promising in providing effective structural priors for compressed sensing, since they can be trained to span low-dimensional data manifolds in high-dimensional signal spaces. Despite the non-convexity of the resulting optimization problem, it has also been shown theoretically that, for neural networks with random Gaussian weights, a signal in the range of the network can be efficiently, approximately recovered from a few noisy measurements. However, a major bottleneck of these theoretical guarantees is a network \emph{expansivity} condition: that each layer of the neural network must be larger than the previous by a logarithmic factor. Our main contribution is to break this strong expansivity assumption, showing that \emph{constant} expansivity suffices to get efficient recovery algorithms, besides it also being information-theoretically necessary. To overcome the theoretical bottleneck in existing approaches we prove a novel uniform concentration theorem for random functions that might not be Lipschitz but satisfy a relaxed notion which we call ``pseudo-Lipschitzness.'' Using this theorem we can show that a matrix concentration inequality known as the \emph{Weight Distribution Condition (WDC)}, which was previously only known to hold for Gaussian matrices with logarithmic aspect ratio, in fact holds for constant aspect ratios too. Since WDC is a fundamental matrix concentration inequality in the heart of all existing theoretical guarantees on this problem, our tighter bound immediately yields improvements in all known results in the literature on compressed sensing with deep generative priors, including one-bit recovery, phase retrieval, and more.

[RANet: Region Attention Network for Semantic Segmentation](#)

- Dingguo Shen, Yuanfeng Ji, Ping Li, Yi Wang, Di Lin
- abstract@[open-review](#): Recent semantic segmentation methods model the relationship between pixels to construct the contextual representations. In this paper, we introduce the \emph{Region Attention Network} (RANet), a novel attention network for modeling the relationship between object regions. RANet divides the image into object regions, where we select representative information. In contrast to the previous methods, RANet configures the information pathways between the pixels in different regions, enabling the region interaction to exchange the regional context for enhancing all of the pixels in the image. We train the construction of object regions, the selection of the representative regional contents, the configuration of information pathways and the context exchange between pixels, jointly, to improve the segmentation accuracy. We extensively evaluate our method on the challenging segmentation benchmarks, demonstrating that RANet effectively helps to achieve the state-of-the-art results.

[A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent](#)

- Zhenyu Liao, Romain Couillet, Michael W. Mahoney
- abstract@[open-review](#): This article characterizes the exact asymptotics of random Fourier feature (RFF) regression, in the realistic setting where the number of data samples n , their dimension p , and the dimension of feature space N are all large and comparable. In this regime, the random RFF Gram matrix no longer converges to the well-known limiting Gaussian kernel matrix (as it does when $N \rightarrow \infty$ alone), but it still has a tractable behavior that is captured by our analysis. This analysis also provides accurate estimates of training and test regression errors for large n, p, N . Based on these estimates, a precise characterization of two qualitatively different phases of learning, including the phase transition between them, is provided; and the corresponding double descent test error curve is derived from this phase transition behavior. These results do not depend on strong assumptions on the data distribution, and they perfectly match empirical results on real-world data sets.

[Learning sparse codes from compressed representations with biologically plausible local wiring constraints](#)

- Kion Fallah, Adam Willats, Ninghao Liu, Christopher Rozell
- abstract@[open-review](#): Sparse coding is an important method for unsupervised learning of task-independent features in theoretical neuroscience models of neural coding. While a number of algorithms exist to learn these representations from the statistics of a dataset, they largely ignore the information bottlenecks present in fiber pathways connecting cortical areas. For example, the visual pathway has many fewer neurons transmitting visual information to cortex than the number of photoreceptors. Both empirical and analytic results have recently shown that sparse representations can be learned effectively after performing dimensionality reduction with randomized linear operators, producing latent coefficients that preserve information. Unfortunately, current proposals for sparse coding in the compressed space require a centralized compression process (i.e., dense random matrix) that is biologically unrealistic due to local wiring constraints observed in neural circuits. The main contribution of this paper is to leverage recent results on structured random matrices to propose a theoretical neuroscience model of randomized projections for communication between cortical areas that is consistent with the local wiring constraints observed in neuroanatomy. We show analytically and empirically that unsupervised learning of sparse representations can be performed in the compressed space despite significant local wiring constraints in compression matrices of varying forms (corresponding to different local wiring patterns). Our analysis verifies that even with significant local wiring constraints, the learned representations remain qualitatively similar, have similar quantitative performance in both training and generalization error, and are consistent across many measures with measured macaque V1 receptive fields.

[Self-Imitation Learning via Generalized Lower Bound Q-learning](#)

- Yunhao Tang
- abstract@[open-review](#): Self-imitation learning motivated by lower-bound Q-learning is a novel and effective approach for off-policy learning. In this work, we propose a n-step lower bound which generalizes the original return-based lower-bound Q-learning, and introduce a new family of self-imitation learning algorithms. To provide a formal motivation for the potential performance gains provided by self-imitation learning, we show that n-step lower bound Q-learning achieves a trade-off between fixed point bias and contraction rate, drawing close connections to the popular uncorrected n-step Q-

learning. We finally show that n-step lower bound Q-learning is a more robust alternative to return-based self-imitation learning and uncorrected n-step, over a wide range of benchmark tasks.

[Private Learning of Halfspaces: Simplifying the Construction and Reducing the Sample Complexity](#)

- Haim Kaplan, Yishay Mansour, Uri Stemmer, Eliad Tsfadia
- abstract@[open-review](#): We present a differentially private learner for halfspaces over a finite grid \mathcal{G} in \mathbb{R}^d with sample complexity $\approx d^{2.5} \cdot 2^{\log^* |\mathcal{G}|}$, which improves the state-of-the-art result of [Beimel et al., COLT 2019] by a d^2 factor. The building block for our learner is a new differentially private algorithm for approximately solving the linear feasibility problem: Given a feasible collection of m linear constraints of the form $Ax \geq b$, the task is to $\{\text{em privately}\}$ identify a solution x that satisfies $\{\text{em most}\}$ of the constraints. Our algorithm is iterative, where each iteration determines the next coordinate of the constructed solution x .

[Directional Pruning of Deep Neural Networks](#)

- Shih-Kang Chao, Zhanyu Wang, Yue Xing, Guang Cheng
- abstract@[open-review](#): In the light of the fact that the stochastic gradient descent (SGD) often finds a flat minimum valley in the training loss, we propose a novel directional pruning method which searches for a sparse minimizer in or close to that flat region. The proposed pruning method does not require retraining or the expert knowledge on the sparsity level. To overcome the computational formidability of estimating the flat directions, we propose to use a carefully tuned ℓ_1 proximal gradient algorithm which can provably achieve the directional pruning with a small learning rate after sufficient training. The empirical results demonstrate the promising results of our solution in highly sparse regime (92% sparsity) among many existing pruning methods on the ResNet50 with the ImageNet, while using only a slightly higher wall time and memory footprint than the SGD. Using the VGG16 and the wide ResNet 28x10 on the CIFAR-10 and CIFAR-100, we demonstrate that our solution reaches the same minima valley as the SGD, and the minima found by our solution and the SGD do not deviate in directions that impact the training loss. The code that reproduces the results of this paper is available at https://github.com/donlan2710/gRDA-Optimizer/tree/master/directional_pruning.

[Smoothly Bounding User Contributions in Differential Privacy](#)

- Alessandro Epasto, Mohammad Mahdian, Jieming Mao, Vahab Mirrokni, Lijie Ren
- abstract@[open-review](#): A differentially private algorithm guarantees that the input of a single user won't significantly change the output distribution of the algorithm. When a user contributes more data points, more information can be collected to improve the algorithm's performance. But at the same time, more noise might need to be added to the algorithm in order to keep the algorithm differentially private and this might hurt the algorithm's performance. Amin et al. (2019) initiates the study on bounding user contributions and proposes a very natural algorithm which limits the number of samples each user can contribute by a threshold. For a better trade-off between utility and privacy guarantee, we propose a method which smoothly bounds user contributions by setting appropriate weights on data points and apply it to estimating the mean/quantiles, linear regression, and empirical risk minimization. We show that our algorithm provably outperforms the sample limiting algorithm. We conclude with experimental evaluations which validate our theoretical results.

[Accelerating Training of Transformer-Based Language Models with Progressive Layer Dropping](#)

- Minjia Zhang, Yuxiong He
- abstract@[open-review](#): Recently, Transformer-based language models have demonstrated remarkable performance across many NLP domains. However, the unsupervised pre-training step of these models suffers from unbearable overall computational expenses. Current methods for accelerating the pre-training either rely on massive parallelism with advanced hardware or are not applicable to language models. In this work, we propose a method based on progressive layer dropping that speeds the training of Transformer-based language models, not at the cost of excessive hardware resources but from model architecture change and training technique boosted efficiency. Extensive experiments on BERT show that the proposed method achieves a 25% reduction of computation cost in FLOPS and a 24% reduction in the end-to-end wall-clock training time. Furthermore, we show that our pre-trained models are equipped with strong knowledge transferability, achieving similar or even higher accuracy in downstream tasks to baseline models.

[Online Planning with Lookahead Policies](#)

- Yonathan Efroni, Mohammad Ghavamzadeh, Shie Mannor
- abstract@[open-review](#): Real Time Dynamic Programming (RTDP) is an online algorithm based on Dynamic Programming (DP) that acts by 1-step greedy planning. Unlike DP, RTDP does not require access to the entire state space, i.e., it explicitly handles the exploration. This fact makes RTDP particularly appealing when the state space is large and it is not possible to update all states simultaneously. In this we devise a multi-step greedy RTDP algorithm, which we call h -RTDP, that replaces the 1-step greedy policy with a h -step lookahead policy. We analyze h -RTDP in its exact form and establish that increasing the lookahead horizon, h , results in an improved sample complexity, with the cost of additional computations. This is the first work that proves improved sample complexity as a result of $\{\text{em increasing}\}$ the lookahead horizon in online planning. We then analyze the performance of h -RTDP in three approximate settings: approximate model, approximate value updates, and approximate state representation. For these cases, we prove that the asymptotic performance of h -RTDP remains the same as that of a corresponding approximate DP algorithm, the best one can hope for without further assumptions on the approximation errors.

[Learning Deep Attribution Priors Based On Prior Knowledge](#)

- Ethan Weinberger, Joseph Janizek, Su-In Lee
- abstract@[open-review](#): Feature attribution methods, which explain an individual prediction made by a model as a sum of attributions for each input feature, are an essential tool for understanding the behavior of complex deep learning models. However, ensuring that models produce meaningful explanations, rather than ones that rely on noise, is not straightforward. Exacerbating this problem is the fact that attribution methods do not provide insight as to why features are assigned their attribution values, leading to explanations that are difficult to interpret. In real-world problems we often have sets of additional information for each feature that are predictive of that feature's importance to the task at hand. Here, we propose the deep attribution prior (DAPr) framework to exploit such information to overcome the limitations of attribution methods. Our framework jointly learns a relationship between prior information and feature importance, as well as biases models to have explanations that rely on features predicted to be important. We find that our framework both results in networks that generalize better to out of sample data and admits new methods for interpreting model behavior.

[Using noise to probe recurrent neural network structure and prune synapses](#)

- Eli Moore, Rishidev Chaudhuri
- abstract@[open-review](#): Many networks in the brain are sparsely connected, and the brain eliminates synapses during development and learning. How could the brain decide which synapses to prune? In a recurrent network, determining the importance of a synapse between two neurons is a difficult computational problem, depending on the role that both neurons play and on all possible pathways of information flow between them. Noise is ubiquitous in neural systems, and often considered an irritant to be overcome. Here we suggest that noise could play a functional role in synaptic pruning, allowing the brain to probe network structure and determine which synapses are redundant. We construct a simple, local, unsupervised plasticity rule that either

strengthens or prunes synapses using only synaptic weight and the noise-driven covariance of the neighboring neurons. For a subset of linear and rectified-linear networks, we prove that this rule preserves the spectrum of the original matrix and hence preserves network dynamics even when the fraction of pruned synapses asymptotically approaches 1. The plasticity rule is biologically-plausible and may suggest a new role for noise in neural computation.

[NanoFlow: Scalable Normalizing Flows with Sublinear Parameter Complexity](#)

- Sang-gil Lee, Sungwon Kim, Sungroh Yoon
- abstract@[open-review](#): Normalizing flows (NFs) have become a prominent method for deep generative models that allow for an analytic probability density estimation and efficient synthesis. However, a flow-based network is considered to be inefficient in parameter complexity because of reduced expressiveness of bijective mapping, which renders the models unfeasibly expensive in terms of parameters. We present an alternative parameterization scheme called NanoFlow, which uses a single neural density estimator to model multiple transformation stages. Hence, we propose an efficient parameter decomposition method and the concept of flow indication embedding, which are key missing components that enable density estimation from a single neural network. Experiments performed on audio and image models confirm that our method provides a new parameter-efficient solution for scalable NFs with significant sublinear parameter complexity.

[Group Knowledge Transfer: Federated Learning of Large CNNs at the Edge](#)

- Chaoyang He, Murali Annavaram, Salman Avestimehr
- abstract@[open-review](#): Scaling up the convolutional neural network (CNN) size (e.g., width, depth, etc.) is known to effectively improve model accuracy. However, the large model size impedes training on resource-constrained edge devices. For instance, federated learning (FL) may place undue burden on the compute capability of edge nodes, even though there is a strong practical need for FL due to its privacy and confidentiality properties. To address the resource-constrained reality of edge devices, we reformulate FL as a group knowledge transfer training algorithm, called FedGKT. FedGKT designs a variant of the alternating minimization approach to train small CNNs on edge nodes and periodically transfer their knowledge by knowledge distillation to a large server-side CNN. FedGKT consolidates several advantages into a single framework: reduced demand for edge computation, lower communication bandwidth for large CNNs, and asynchronous training, all while maintaining model accuracy comparable to FedAvg. We train CNNs designed based on ResNet-56 and ResNet-110 using three distinct datasets (CIFAR-10, CIFAR-100, and CINIC-10) and their non-IID variants. Our results show that FedGKT can obtain comparable or even slightly higher accuracy than FedAvg. More importantly, FedGKT makes edge training affordable. Compared to the edge training using FedAvg, FedGKT demands 9 to 17 times less computational power (FLOPs) on edge devices and requires 54 to 105 times fewer parameters in the edge CNN. Our source code is released at FedML (<https://fedml.ai>).

[Neural FFTs for Universal Texture Image Synthesis](#)

- Morteza Mardani, Guilin Liu, Aysegul Dundar, Shiqiu Liu, Andrew Tao, Bryan Catanzaro
- abstract@[open-review](#): Synthesizing larger texture images from a smaller exemplar is an important task in graphics and vision. The conventional CNNs, recently adopted for synthesis, require to train and test on the same set of images and fail to generalize to unseen images. This is mainly because those CNNs fully rely on convolutional and upsampling layers that operate locally and not suitable for a task as global as texture synthesis. In this work, inspired by the repetitive nature of texture patterns, we find that texture synthesis can be viewed as (local) \textit{upsampling} in the Fast Fourier Transform (FFT) domain. However, FFT of natural images exhibits high dynamic range and lacks local correlations. Therefore, to train CNNs we design a framework to perform FFT upsampling in feature space using deformable convolutions. Such design allows our framework to generalize to unseen images, and synthesize textures in a single pass. Extensive evaluations confirm that our method achieves state-of-the-art performance both quantitatively and qualitatively.

[Graph Cross Networks with Vertex Infomax Pooling](#)

- Maosen Li, Siheng Chen, Ya Zhang, Ivor Tsang
- abstract@[open-review](#): We propose a novel graph cross network (GXN) to achieve comprehensive feature learning from multiple scales of a graph. Based on trainable hierarchical representations of a graph, GXN enables the interchange of intermediate features across scales to promote information flow. Two key ingredients of GXN include a novel vertex infomax pooling (VIPool), which creates multiscale graphs in a trainable manner, and a novel feature-crossing layer, enabling feature interchange across scales. The proposed VIPool selects the most informative subset of vertices based on the neural estimation of mutual information between vertex features and neighborhood features. The intuition behind is that a vertex is informative when it can maximally reflect its neighboring information. The proposed feature-crossing layer fuses intermediate features between two scales for mutual enhancement by improving information flow and enriching multiscale features at hidden layers. The cross shape of feature-crossing layer distinguishes GXN from many other multiscale architectures. Experimental results show that the proposed GXN improves the classification accuracy by 2.12% and 1.15% on average for graph classification and vertex classification, respectively. Based on the same network, the proposed VIPool consistently outperforms other graph-pooling methods.

[Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms](#)

- Hilal Asi, John C. Duchi
- abstract@[open-review](#): We study and provide instance-optimal algorithms in differential privacy by extending and approximating the inverse sensitivity mechanism. We provide two approximation frameworks, one which only requires knowledge of local sensitivities, and a gradient-based approximation for optimization problems, which are efficiently computable for a broad class of functions. We complement our analysis with instance-specific lower bounds for vector-valued functions, which demonstrate that our mechanisms are (nearly) instance-optimal under certain assumptions and that minimax lower bounds may not provide an accurate estimate of the hardness of a problem in general: our algorithms can significantly outperform minimax bounds for well-behaved instances. Finally, we use our approximation framework to develop private mechanisms for unbounded-range mean estimation, principal component analysis, and linear regression. For PCA, our mechanisms give an efficient (pure) differentially private algorithm with near-optimal rates.

[Calibration of Shared Equilibria in General Sum Partially Observable Markov Games](#)

- Nelson Vadovi, Sumitra Ganesh, Prashant Reddy, Manuela Veloso
- abstract@[open-review](#): Training multi-agent systems (MAS) to achieve realistic equilibria gives us a useful tool to understand and model real-world systems. We consider a general sum partially observable Markov game where agents of different types share a single policy network, conditioned on agent-specific information. This paper aims at i) formally understanding equilibria reached by such agents, and ii) matching emergent phenomena of such equilibria to real-world targets. Parameter sharing with decentralized execution has been introduced as an efficient way to train multiple agents using a single policy network. However, the nature of resulting equilibria reached by such agents has not been yet studied: we introduce the novel concept of Shared equilibrium as a symmetric pure Nash equilibrium of a certain Functional Form Game (FFG) and prove convergence to the latter for a certain class of games using self-play. In addition, it is important that such equilibria satisfy certain constraints so that MAS are calibrated to real world data for practical use: we solve this problem by introducing a novel dual-Reinforcement Learning based approach that fits emergent behaviors of agents in a Shared equilibrium to externally-specified targets, and apply our methods to a n-player market example. We do so by calibrating parameters governing distributions of agent types rather than individual agents, which allows both behavior differentiation among agents and coherent scaling of the shared policy network to multiple agents.

MOPO: Model-based Offline Policy Optimization

- Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y. Zou, Sergey Levine, Chelsea Finn, Tengyu Ma
- abstract@[open-review](#): Offline reinforcement learning (RL) refers to the problem of learning policies entirely from a batch of previously collected data. This problem setting is compelling, because it offers the promise of utilizing large, diverse, previously collected datasets to acquire policies without any costly or dangerous active exploration, but it is also exceptionally difficult, due to the distributional shift between the offline training data and the learned policy. While there has been significant progress in model-free offline RL, the most successful prior methods constrain the policy to the support of the data, precluding generalization to new states. In this paper, we observe that an existing model-based RL algorithm on its own already produces significant gains in the offline setting, as compared to model-free approaches, despite not being designed for this setting. However, although many standard model-based RL methods already estimate the uncertainty of their model, they do not by themselves provide a mechanism to avoid the issues associated with distributional shift in the offline setting. We therefore propose to modify existing model-based RL methods to address these issues by casting offline model-based RL into a penalized MDP framework. We theoretically show that, by using this penalized MDP, we are maximizing a lower bound of the return in the true MDP. Based on our theoretical results, we propose a new model-based offline RL algorithm that applies the variance of a Lipschitz-regularized model as a penalty to the reward function. We find that this algorithm outperforms both standard model-based RL methods and existing state-of-the-art model-free offline RL approaches on existing offline RL benchmarks, as well as two challenging continuous control tasks that require generalizing from data collected for a different task.

Building powerful and equivariant graph neural networks with structural message-passing

- Clément Vignac, Andreas Loukas, Pascal Frossard
- abstract@[open-review](#): Message-passing has proved to be an effective way to design graph neural networks, as it is able to leverage both permutation equivariance and an inductive bias towards learning local structures in order to achieve good generalization. However, current message-passing architectures have a limited representation power and fail to learn basic topological properties of graphs. We address this problem and propose a powerful and equivariant message-passing framework based on two ideas: first, we propagate a one-hot encoding of the nodes, in addition to the features, in order to learn a local context matrix around each node. This matrix contains rich local information about both features and topology and can eventually be pooled to build node representations. Second, we propose methods for the parametrization of the message and update functions that ensure permutation equivariance. Having a representation that is independent of the specific choice of the one-hot encoding permits inductive reasoning and leads to better generalization properties. Experimentally, our model can predict various graph topological properties on synthetic data more accurately than previous methods and achieves state-of-the-art results on molecular graph regression on the ZINC dataset.

Efficient Model-Based Reinforcement Learning through Optimistic Policy Search and Planning

- Sebastian Curi, Felix Berkenkamp, Andreas Krause
- abstract@[open-review](#): Model-based reinforcement learning algorithms with probabilistic dynamical models are amongst the most data-efficient learning methods. This is often attributed to their ability to distinguish between epistemic and aleatoric uncertainty. However, while most algorithms distinguish these two uncertainties for learning the model, they ignore it when optimizing the policy, which leads to greedy and insufficient exploration. At the same time, there are no practical solvers for optimistic exploration algorithms. In this paper, we propose a practical optimistic exploration algorithm (H-UCRL). H-UCRL reparameterizes the set of plausible models and hallucinates control directly on the epistemic uncertainty. By augmenting the input space with the hallucinated inputs, H-UCRL can be solved using standard greedy planners. Furthermore, we analyze H-UCRL and construct a general regret bound for well-calibrated models, which is provably sublinear in the case of Gaussian Process models. Based on this theoretical foundation, we show how optimistic exploration can be easily combined with state-of-the-art reinforcement learning algorithms and different probabilistic models. Our experiments demonstrate that optimistic exploration significantly speeds-up learning when there are penalties on actions, a setting that is notoriously difficult for existing model-based reinforcement learning algorithms.

Practical Low-Rank Communication Compression in Decentralized Deep Learning

- Thijs Vogels, Sai Praneeth Karimireddy, Martin Jaggi
- abstract@[open-review](#): Lossy gradient compression has become a practical tool to overcome the communication bottleneck in centrally coordinated distributed training of machine learning models. However, algorithms for decentralized training with compressed communication over arbitrary connected networks have been more complicated, requiring additional memory and hyperparameters. We introduce a simple algorithm that directly compresses the model differences between neighboring workers using low-rank linear compressors. We prove that our method does not require any additional hyperparameters, converges faster than prior methods, and is asymptotically independent of both the network and the compression. Inspired by the PowerSGD algorithm for centralized deep learning, we execute power iteration steps on model differences to maximize the information transferred per bit. Out of the box, these compressors perform on par with state-of-the-art tuned compression algorithms in a series of deep learning benchmarks.

Mutual exclusivity as a challenge for deep neural networks

- Kanishk Gandhi, Brenden M. Lake
- abstract@[open-review](#): Strong inductive biases allow children to learn in fast and adaptable ways. Children use the mutual exclusivity (ME) bias to help disambiguate how words map to referents, assuming that if an object has one label then it does not need another. In this paper, we investigate whether or not vanilla neural architectures have an ME bias, demonstrating that they lack this learning assumption. Moreover, we show that their inductive biases are poorly matched to lifelong learning formulations of classification and translation. We demonstrate that there is a compelling case for designing task-general neural networks that learn through mutual exclusivity, which remains an open challenge.

3D Shape Reconstruction from Vision and Touch

- Edward Smith, Roberto Calandra, Adriana Romero, Georgia Gkioxari, David Meger, Jitendra Malik, Michal Drozdzal
- abstract@[open-review](#): When a toddler is presented a new toy, their instinctual behaviour is to pick it up and inspect it with their hand and eyes in tandem, clearly searching over its surface to properly understand what they are playing with. At any instance here, touch provides high fidelity localized information while vision provides complementary global context. However, in 3D shape reconstruction, the complementary fusion of visual and haptic modalities remains largely unexplored. In this paper, we study this problem and present an effective chart-based approach to multi-modal shape understanding which encourages a similar fusion of vision and touch information. To do so, we introduce a dataset of simulated touch and vision signals from the interaction between a robotic hand and a large array of 3D objects. Our results show that (1) leveraging both vision and touch signals consistently improves single-modality baselines; (2) our approach outperforms alternative modality fusion methods and strongly benefits from the proposed chart-based structure; (3) the reconstruction quality increases with the number of grasps provided; and (4) the touch information not only enhances the reconstruction at the touch site but also extrapolates to its local neighborhood.

GradAug: A New Regularization Method for Deep Neural Networks

- Taojannan Yang, Sijie Zhu, Chen Chen

- abstract@[open-review](#): We propose a new regularization method to alleviate over-fitting in deep neural networks. The key idea is utilizing randomly transformed training samples to regularize a set of sub-networks, which are originated by sampling the width of the original network, in the training process. As such, the proposed method introduces self-guided disturbances to the raw gradients of the network and therefore is termed as Gradient Augmentation (GradAug). We demonstrate that GradAug can help the network learn well-generalized and more diverse representations. Moreover, it is easy to implement and can be applied to various structures and applications. GradAug improves ResNet-50 to 78.79% on ImageNet classification, which is a new state-of-the-art accuracy. By combining with CutMix, it further boosts the performance to 79.67%, which outperforms an ensemble of advanced training tricks. The generalization ability is evaluated on COCO object detection and instance segmentation where GradAug significantly surpasses other state-of-the-art methods. GradAug is also robust to image distortions and FGSM adversarial attacks and is highly effective in low data regimes. Code is available at \url{https://github.com/taoyang1122/GradAug}

[An Equivalence between Loss Functions and Non-Uniform Sampling in Experience Replay](#)

- Scott Fujimoto, David Meger, Doina Precup
- abstract@[open-review](#): Prioritized Experience Replay (PER) is a deep reinforcement learning technique in which agents learn from transitions sampled with non-uniform probability proportionate to their temporal-difference error. We show that any loss function evaluated with non-uniformly sampled data can be transformed into another uniformly sampled loss function with the same expected gradient. Surprisingly, we find in some environments PER can be replaced entirely by this new loss function without impact to empirical performance. Furthermore, this relationship suggests a new branch of improvements to PER by correcting its uniformly sampled loss function equivalent. We demonstrate the effectiveness of our proposed modifications to PER and the equivalent loss function in several MuJoCo and Atari environments.

[Learning Utilities and Equilibria in Non-Truthful Auctions](#)

- Hu Fu, Tao Lin
- abstract@[open-review](#): In non-truthful auctions, agents' utility for a strategy depends on the strategies of the opponents and also the prior distribution over their private types; the set of Bayes Nash equilibria generally has an intricate dependence on the prior. Using the First Price Auction as our main demonstrating example, we show that $\tilde{O}(n / \epsilon^2)$ samples from the prior with n agents suffice for an algorithm to learn the interim utilities for all monotone bidding strategies. As a consequence, this number of samples suffice for learning all approximate equilibria. We give almost matching (up to polylog factors) lower bound on the sample complexity for learning utilities. We also consider a setting where agents must pay a search cost to discover their own types. Drawing on a connection between this setting and the first price auction, discovered recently by Kleinberg et al. (2016), we show that $\tilde{O}(n / \epsilon^2)$ samples suffice for utilities and equilibria to be estimated in a near welfare-optimal descending auction in this setting. En route, we improve the sample complexity bound, recently obtained by Guo et al. (2019), for the Pandora's Box problem, which is a classical model for sequential consumer search.

[Rational neural networks](#)

- Nicolas Boule, Yuji Nakatsukasa, Alex Townsend
- abstract@[open-review](#): We consider neural networks with rational activation functions. The choice of the nonlinear activation function in deep learning architectures is crucial and heavily impacts the performance of a neural network. We establish optimal bounds in terms of network complexity and prove that rational neural networks approximate smooth functions more efficiently than ReLU networks with exponentially smaller depth. The flexibility and smoothness of rational activation functions make them an attractive alternative to ReLU, as we demonstrate with numerical experiments.

[DISK: Learning local features with policy gradient](#)

- MichaÅ, Tyszkiewicz, Pascal Fua, Eduard Trulls
- abstract@[open-review](#): Local feature frameworks are difficult to learn in an end-to-end fashion due to the discreteness inherent to the selection and matching of sparse keypoints. We introduce DISK (DIScrete Keypoints), a novel method that overcomes these obstacles by leveraging principles from Reinforcement Learning (RL), optimizing end-to-end for a high number of correct feature matches. Our simple yet expressive probabilistic model lets us keep the training and inference regimes close, while maintaining good enough convergence properties to reliably train from scratch. Our features can be extracted very densely while remaining discriminative, challenging commonly held assumptions about what constitutes a good keypoint, as showcased in Fig. 1, and deliver state-of-the-art results on three public benchmarks.

[Transfer Learning via \$\ell_1\$ Regularization](#)

- Masaaki Takada, Hironori Fujisawa
- abstract@[open-review](#): Machine learning algorithms typically require abundant data under a stationary environment. However, environments are nonstationary in many real-world applications. Critical issues lie in how to effectively adapt models under an ever-changing environment. We propose a method for transferring knowledge from a source domain to a target domain via ℓ_1 regularization in high dimension. We incorporate ℓ_1 regularization of differences between source and target parameters in addition to an ordinary ℓ_1 regularization. Hence, our method yields sparsity for both the estimates themselves and changes of the estimates. The proposed method has a tight estimation error bound under a stationary environment, and the estimate remains unchanged from the source estimate under small residuals. Moreover, the estimate is consistent with the underlying function, even when the source estimate is mistaken due to nonstationarity. Empirical results demonstrate that the proposed method effectively balances stability and plasticity.

[GOCor: Bringing Globally Optimized Correspondence Volumes into Your Neural Network](#)

- Prune Truong, Martin Danelljan, Luc V. Gool, Radu Timofte
- abstract@[open-review](#): The feature correlation layer serves as a key neural network module in numerous computer vision problems that involve dense correspondences between image pairs. It predicts a correspondence volume by evaluating dense scalar products between feature vectors extracted from pairs of locations in two images. However, this point-to-point feature comparison is insufficient when disambiguating multiple similar regions in an image, severely affecting the performance of the end task. We propose GOCor, a fully differentiable dense matching module, acting as a direct replacement to the feature correlation layer. The correspondence volume generated by our module is the result of an internal optimization procedure that explicitly accounts for similar regions in the scene. Moreover, our approach is capable of effectively learning spatial matching priors to resolve further matching ambiguities. We analyze our GOCor module in extensive ablative experiments. When integrated into state-of-the-art networks, our approach significantly outperforms the feature correlation layer for the tasks of geometric matching, optical flow, and dense semantic matching. The code and trained models will be made available at github.com/PruneTruong/GOCor.

[Deep Inverse Q-learning with Constraints](#)

- Gabriel Kalweit, Maria Huegle, Moritz Werling, Joschka Boedecker
- abstract@[open-review](#): Popular Maximum Entropy Inverse Reinforcement Learning approaches require the computation of expected state visitation frequencies for the optimal policy under an estimate of the reward function. This usually requires intermediate value estimation in the inner loop of the

algorithm, slowing down convergence considerably. In this work, we introduce a novel class of algorithms that only needs to solve the MDP underlying the demonstrated behavior once to recover the expert policy. This is possible through a formulation that exploits a probabilistic behavior assumption for the demonstrations within the structure of Q-learning. We propose Inverse Action-value Iteration which is able to fully recover an underlying reward of an external agent in closed-form analytically. We further provide an accompanying class of sampling-based variants which do not depend on a model of the environment. We show how to extend this class of algorithms to continuous state-spaces via function approximation and how to estimate a corresponding action-value function, leading to a policy as close as possible to the policy of the external agent, while optionally satisfying a list of predefined hard constraints. We evaluate the resulting algorithms called Inverse Action-value Iteration, Inverse Q-learning and Deep Inverse Q-learning on the Objectworld benchmark, showing a speedup of up to several orders of magnitude compared to (Deep) Max-Entropy algorithms. We further apply Deep Constrained Inverse Q-learning on the task of learning autonomous lane-changes in the open-source simulator SUMO achieving competent driving after training on data corresponding to 30 minutes of demonstrations.

[Optimistic Dual Extrapolation for Coherent Non-monotone Variational Inequalities](#)

- Chaobing Song, Zhengyuan Zhou, Yichao Zhou, Yong Jiang, Yi Ma
- abstract@[open-review](#): The optimization problems associated with training generative adversarial neural networks can be largely reduced to certain {\em non-monotone} variational inequality problems (VIPs), whereas existing convergence results are mostly based on monotone or strongly monotone assumptions. In this paper, we propose {\em optimistic dual extrapolation (OptDE)}, a method that only performs {\em one} gradient evaluation per iteration. We show that OptDE is provably convergent to {\em a strong solution} under different coherent non-monotone assumptions. In particular, when a {\em weak solution} exists, the convergence rate of our method is $\mathcal{O}(1/\{\epsilon\}^2)$, which matches the best existing result of the methods with two gradient evaluations. Further, when a {\em \$\sigma\$-weak solution} exists, the convergence guarantee is improved to the linear rate $\mathcal{O}(\log\frac{1}{\epsilon})$. Along the way--as a byproduct of our inquiries into non-monotone variational inequalities--we provide the near-optimal $\mathcal{O}(\log\frac{1}{\epsilon})$ convergence guarantee in terms of restricted strong merit function for monotone variational inequalities. We also show how our results can be naturally generalized to the stochastic setting, and obtain corresponding new convergence results. Taken together, our results contribute to the broad landscape of variational inequality--both non-monotone and monotone alike--by providing a novel and more practical algorithm with the state-of-the-art convergence guarantees.

[Prediction with Corrupted Expert Advice](#)

- Idan Amir, Idan Attias, Tomer Koren, Yishay Mansour, Roi Livni
- abstract@[open-review](#): We revisit the fundamental problem of prediction with expert advice, in a setting where the environment is benign and generates losses stochastically, but the feedback observed by the learner is subject to a moderate adversarial corruption. We prove that a variant of the classical Multiplicative Weights algorithm with decreasing step sizes achieves constant regret in this setting and performs optimally in a wide range of environments, regardless of the magnitude of the injected corruption. Our results reveal a surprising disparity between the often comparable Follow the Regularized Leader (FTRL) and Online Mirror Descent (OMD) frameworks: we show that for experts in the corrupted stochastic regime, the regret performance of OMD is in fact strictly inferior to that of FTRL.

[Human Parsing Based Texture Transfer from Single Image to 3D Human via Cross-View Consistency](#)

- Fang Zhao, Shengcai Liao, Kaihao Zhang, Ling Shao
- abstract@[open-review](#): This paper proposes a human parsing based texture transfer model via cross-view consistency learning to generate the texture of 3D human body from a single image. We use the semantic parsing of human body as input for providing both the shape and pose information to reduce the appearance variation of human image and preserve the spatial distribution of semantic parts. Meanwhile, in order to improve the prediction for textures of invisible parts, we explicitly enforce the consistency across different views of the same subject by exchanging the textures predicted by two views to render images during training. The perception loss and total variation regularization are optimized to maximize the similarity between rendered and input images, which does not necessitate extra 3D texture supervision. Experimental results on pedestrian images and fashion photos demonstrate that our method can produce higher quality textures with convincing details than other texture generation methods.

[Knowledge Augmented Deep Neural Networks for Joint Facial Expression and Action Unit Recognition](#)

- Zijun Cui, Tengfei Song, Yuru Wang, Qiang Ji
- abstract@[open-review](#): Facial expression and action units (AUs) represent two levels of descriptions of the facial behavior. Due to the underlying facial anatomy and the need to form a meaningful coherent expression, they are strongly correlated. This paper proposes to systematically capture their dependencies and incorporate them into a deep learning framework for joint facial expression recognition and action unit detection. Specifically, we first propose a constraint optimization method to encode the generic knowledge on expression-AUs probabilistic dependencies into a Bayesian Network (BN). The BN is then integrated into a deep learning framework as a weak supervision for an AU detection model. A data-driven facial expression recognition(FER) model is then constructed from data. Finally, the FER model and AU detection model are trained jointly to refine their learning. Evaluations on benchmark datasets demonstrate the effectiveness of the proposed knowledge integration in improving the performance of both the FER model and the AU detection model. The proposed AU detection model is demonstrated to be able to achieve competitive performance without AU annotations. Furthermore, the proposed Bayesian Network capturing the generic knowledge is demonstrated to generalize well to different datasets.

[Point process models for sequence detection in high-dimensional neural spike trains](#)

- Alex Williams, Anthony Degleris, Yixin Wang, Scott Linderman
- abstract@[open-review](#): Sparse sequences of neural spikes are posited to underlie aspects of working memory, motor production, and learning. Discovering these sequences in an unsupervised manner is a longstanding problem in statistical neuroscience. Promising recent work utilized a convolutive nonnegative matrix factorization model to tackle this challenge. However, this model requires spike times to be discretized, utilizes a sub-optimal least-squares criterion, and does not provide uncertainty estimates for model predictions or estimated parameters. We address each of these shortcomings by developing a point process model that characterizes fine-scale sequences at the level of individual spikes and represents sequence occurrences as a small number of marked events in continuous time. This ultra-sparse representation of sequence events opens new possibilities for spike train modeling. For example, we introduce learnable time warping parameters to model sequences of varying duration, which have been experimentally observed in neural circuits. We demonstrate these advantages on recordings from songbird higher vocal center and rodent hippocampus.

[Adversarial Attacks on Linear Contextual Bandits](#)

- Evrard Garcelon, Baptiste Roziere, Laurent Meunier, Jean Tarbouriech, Olivier Teytaud, Alessandro Lazaric, Matteo Pirotta
- abstract@[open-review](#): Contextual bandit algorithms are applied in a wide range of domains, from advertising to recommender systems, from clinical trials to education. In many of these domains, malicious agents may have incentives to force a bandit algorithm into a desired behavior. For instance, an unscrupulous ad publisher may try to increase their own revenue at the expense of the advertisers; a seller may want to increase the exposure of their products, or thwart a competitor's advertising campaign. In this paper, we study several attack scenarios and show that a malicious agent can force a linear contextual bandit algorithm to pull any desired arm $T \geq \mathcal{O}(T)$ times over a horizon of T steps, while applying adversarial modifications to either rewards or contexts with a cumulative cost that only grow logarithmically as $\mathcal{O}(\log T)$. We also investigate the case when a malicious agent is interested in

affecting the behavior of the bandit algorithm in a single context (e.g., a specific user). We first provide sufficient conditions for the feasibility of the attack and an efficient algorithm to perform an attack. We empirically validate the proposed approaches on synthetic and real-world datasets.

[Meta-Consolidation for Continual Learning](#)

- Joseph K J, Vineeth N Balasubramanian
- abstract@[open-review](#): The ability to continuously learn and adapt itself to new tasks, without losing grasp of already acquired knowledge is a hallmark of biological learning systems, which current deep learning systems fall short of. In this work, we present a novel methodology for continual learning called MERLIN: Meta-Consolidation for Continual Learning. We assume that weights of a neural network, for solving task, come from a meta-distribution. This meta-distribution is learned and consolidated incrementally. We operate in the challenging online continual learning setting, where a data point is seen by the model only once. Our experiments with continual learning benchmarks of MNIST, CIFAR-10, CIFAR-100 and Mini-ImageNet datasets show consistent improvement over five baselines, including a recent state-of-the-art, corroborating the promise of MERLIN.

[Organizing recurrent network dynamics by task-computation to enable continual learning](#)

- Lea Duncker, Laura Driscoll, Krishna V. Shenoy, Maneesh Sahani, David Sussillo
- abstract@[open-review](#): Biological systems face dynamic environments that require continual learning. It is not well understood how these systems balance the tension between flexibility for learning and robustness for memory of previous behaviors. Continual learning without catastrophic interference also remains a challenging problem in machine learning. Here, we develop a novel learning rule designed to minimize interference between sequentially learned tasks in recurrent networks. Our learning rule preserves network dynamics within activity-defined subspaces used for previously learned tasks. It encourages dynamics associated with new tasks that might otherwise interfere to instead explore orthogonal subspaces, and it allows for reuse of previously established dynamical motifs where possible. Employing a set of tasks used in neuroscience, we demonstrate that our approach successfully eliminates catastrophic interference and offers a substantial improvement over previous continual learning algorithms. Using dynamical systems analysis, we show that networks trained using our approach can reuse similar dynamical structures across similar tasks. This possibility for shared computation allows for faster learning during sequential training. Finally, we identify organizational differences that emerge when training tasks sequentially versus simultaneously.

[Lifelong Policy Gradient Learning of Factored Policies for Faster Training Without Forgetting](#)

- Jorge Mendez, Boyu Wang, Eric Eaton
- abstract@[open-review](#): Policy gradient methods have shown success in learning control policies for high-dimensional dynamical systems. Their biggest downside is the amount of exploration they require before yielding high-performing policies. In a lifelong learning setting, in which an agent is faced with multiple consecutive tasks over its lifetime, reusing information from previously seen tasks can substantially accelerate the learning of new tasks. We provide a novel method for lifelong policy gradient learning that trains lifelong function approximators directly via policy gradients, allowing the agent to benefit from accumulated knowledge throughout the entire training process. We show empirically that our algorithm learns faster and converges to better policies than single-task and lifelong learning baselines, and completely avoids catastrophic forgetting on a variety of challenging domains.

[Kernel Methods Through the Roof: Handling Billions of Points Efficiently](#)

- Giacomo Meanti, Luigi Carratino, Lorenzo Rosasco, Alessandro Rudi
- abstract@[open-review](#): Kernel methods provide an elegant and principled approach to nonparametric learning, but so far could hardly be used in large scale problems, since naïve implementations scale poorly with data size. Recent advances have shown the benefits of a number of algorithmic ideas, for example combining optimization, numerical linear algebra and random projections. Here, we push these efforts further to develop and test a solver that takes full advantage of GPU hardware. Towards this end, we designed a preconditioned gradient solver for kernel methods exploiting both GPU acceleration and parallelization with multiple GPUs, implementing out-of-core variants of common linear algebra operations to guarantee optimal hardware utilization. Further, we optimize the numerical precision of different operations and maximize efficiency of matrix-vector multiplications. As a result we can experimentally show dramatic speedups on datasets with billions of points, while still guaranteeing state of the art performance. Additionally, we make our software available as an easy to use library.

[Spike and slab variational Bayes for high dimensional logistic regression](#)

- Kolyan Ray, Botond Szabo, Gabriel Clara
- abstract@[open-review](#): Variational Bayes (VB) is a popular scalable alternative to Markov chain Monte Carlo for Bayesian inference. We study a mean-field spike and slab VB approximation of widely used Bayesian model selection priors in sparse high-dimensional logistic regression. We provide non-asymptotic theoretical guarantees for the VB posterior in both ℓ_2 and prediction loss for a sparse truth, giving optimal (minimax) convergence rates. Since the VB algorithm does not depend on the unknown truth to achieve optimality, our results shed light on effective prior choices. We confirm the improved performance of our VB algorithm over common sparse VB approaches in a numerical study.

[Maximum-Entropy Adversarial Data Augmentation for Improved Generalization and Robustness](#)

- Long Zhao, Ting Liu, Xi Peng, Dimitris Metaxas
- abstract@[open-review](#): Adversarial data augmentation has shown promise for training robust deep neural networks against unforeseen data shifts or corruptions. However, it is difficult to define heuristics to generate effective fictitious target distributions containing "hard" adversarial perturbations that are largely different from the source distribution. In this paper, we propose a novel and effective regularization term for adversarial data augmentation. We theoretically derive it from the information bottleneck principle, which results in a maximum-entropy formulation. Intuitively, this regularization term encourages perturbing the underlying source distribution to enlarge predictive uncertainty of the current model, so that the generated "hard" adversarial perturbations can improve the model robustness during training. Experimental results on three standard benchmarks demonstrate that our method consistently outperforms the existing state of the art by a statistically significant margin.

[Fast geometric learning with symbolic matrices](#)

- Jean Feydy, Alexis Glauñ's, Benjamin Charlier, Michael Bronstein
- abstract@[open-review](#): Geometric methods rely on tensors that can be encoded using a symbolic formula and data arrays, such as kernel and distance matrices. We present an extension for standard machine learning frameworks that provides comprehensive support for this abstraction on CPUs and GPUs: our toolbox combines a versatile, transparent user interface with fast runtimes and low memory usage. Unlike general purpose acceleration frameworks such as XLA, our library turns generic Python code into binaries whose performances are competitive with state-of-the-art geometric libraries - such as FAISS for nearest neighbor search - with the added benefit of flexibility. We perform an extensive evaluation on a broad class of problems: Gaussian modelling, K-nearest neighbors search, geometric deep learning, non-Euclidean embeddings and optimal transport theory. In practice, for geometric problems that involve 1k to 1M samples in dimension 1 to 100, our library speeds up baseline GPU implementations by up to two orders of magnitude.

[MESA: Boost Ensemble Imbalanced Learning with MEda-SAmpler](#)

- Zhining Liu, Pengfei Wei, Jing Jiang, Wei Cao, Jiang Bian, Yi Chang
- abstract@[open-review](#): Imbalanced learning (IL), i.e., learning unbiased models from class-imbalanced data, is a challenging problem. Typical IL methods including resampling and reweighting were designed based on some heuristic assumptions. They often suffer from unstable performance, poor applicability, and high computational cost in complex tasks where their assumptions do not hold. In this paper, we introduce a novel ensemble IL framework named MESA. It adaptively resamples the training set in iterations to get multiple classifiers and forms a cascade ensemble model. MESA directly learns the sampling strategy from data to optimize the final metric beyond following random heuristics. Moreover, unlike prevailing meta-learning-based IL solutions, we decouple the model-training and meta-training in MESA by independently train the meta-sampler over task-agnostic meta-data. This makes MESA generally applicable to most of the existing learning models and the meta-sampler can be efficiently applied to new tasks. Extensive experiments on both synthetic and real-world tasks demonstrate the effectiveness, robustness, and transferability of MESA. Our code is available at <https://github.com/ZhiningLiu1998/mesa>.

[CoinPress: Practical Private Mean and Covariance Estimation](#)

- Sourav Biswas, Yihe Dong, Gautam Kamath, Jonathan Ullman
- abstract@[open-review](#): We present simple differentially private estimators for the parameters of multivariate sub-Gaussian data that are accurate at small sample sizes. We demonstrate the effectiveness of our algorithms both theoretically and empirically using synthetic and real-world datasets---showing that their asymptotic error rates match the state-of-the-art theoretical bounds, and that they concretely outperform all previous methods. Specifically, previous estimators either have weak empirical accuracy at small sample sizes, perform poorly for multivariate data, or require the user to provide strong a priori estimates for the parameters.

[Planning with General Objective Functions: Going Beyond Total Rewards](#)

- Ruosong Wang, Peilin Zhong, Simon S. Du, Russ R. Salakhutdinov, Lin Yang
- abstract@[open-review](#): Standard sequential decision-making paradigms aim to maximize the cumulative reward when interacting with the unknown environment., i.e., maximize $\sum_{h=1}^H r_h$ where H is the planning horizon. However, this paradigm fails to model important practical applications, e.g., safe control that aims to maximize the lowest reward, i.e., maximize $\min_{h=1}^H r_h$. In this paper, based on techniques in sketching algorithms, we propose a novel planning algorithm in deterministic systems which deals with a large class of objective functions of the form $f(r_1, r_2, \dots, r_H)$ that are of interest to practical applications. We show that efficient planning is possible if f is symmetric under permutation of coordinates and satisfies certain technical conditions. Complementing our algorithm, we further prove that removing any of the conditions will make the problem intractable in the worst case and thus demonstrate the necessity of our conditions.

[Scattering GCN: Overcoming Oversmoothness in Graph Convolutional Networks](#)

- Yimeng Min, Frederik Wenkel, Guy Wolf
- abstract@[open-review](#): Graph convolutional networks (GCNs) have shown promising results in processing graph data by extracting structure-aware features. This gave rise to extensive work in geometric deep learning, focusing on designing network architectures that ensure neuron activations conform to regularity patterns within the input graph. However, in most cases the graph structure is only accounted for by considering the similarity of activations between adjacent nodes, which limits the capabilities of such methods to discriminate between nodes in a graph. Here, we propose to augment conventional GCNs with geometric scattering transforms and residual convolutions. The former enables band-pass filtering of graph signals, thus alleviating the so-called oversmoothing often encountered in GCNs, while the latter is introduced to clear the resulting features of high-frequency noise. We establish the advantages of the presented Scattering GCN with both theoretical results establishing the complementary benefits of scattering and GCN features, as well as experimental results showing the benefits of our method compared to leading graph neural networks for semi-supervised node classification, including the recently proposed GAT network that typically alleviates oversmoothing using graph attention mechanisms.

[KFC: A Scalable Approximation Algorithm for \$k\$ -center Fair Clustering](#)

- Elfarouk Harb, Ho Shan Lam
- abstract@[open-review](#): In this paper, we study the problem of fair clustering on the k -center objective. In fair clustering, the input is N points, each belonging to at least one of I protected groups, e.g. male, female, Asian, Hispanic. The objective is to cluster the N points into k clusters to minimize a classical clustering objective function. However, there is an additional constraint that each cluster needs to be fair, under some notion of fairness. This ensures that no group is either "over-represented" or "under-represented" in any cluster. Our work builds on the work of Chierichetti et al. (NIPS 2017), Bera et al. (NeurIPS 2019), Ahmadian et al. (KDD 2019), and Bercea et al. (APPROX 2019). We obtain a randomized 3 -approximation algorithm for the k -center objective function, beating the previous state of the art (4 -approximation). We test our algorithm on real datasets, and show that our algorithm is effective in finding good clusters without over-representation or under-representation, surpassing the current state of the art in runtime speed, clustering cost, while achieving similar fairness violations.

[Leveraging Predictions in Smoothed Online Convex Optimization via Gradient-based Algorithms](#)

- Yingying Li, Na Li
- abstract@[open-review](#): We consider online convex optimization with time-varying stage costs and additional switching costs. Since the switching costs introduce coupling across all stages, multi-step-ahead (long-term) predictions are incorporated to improve the online performance. However, longer-term predictions tend to suffer from lower quality. Thus, a critical question is: how to reduce the impact of long-term prediction errors on the online performance? To address this question, we introduce a gradient-based online algorithm, Receding Horizon Inexact Gradient (RHIG), and analyze its performance by dynamic regrets in terms of the temporal variation of the environment and the prediction errors. RHIG only considers at most W -step-ahead predictions to avoid being misled by worse predictions in the longer term. The optimal choice of W suggested by our regret bounds depends on the tradeoff between the variation of the environment and the prediction accuracy. Additionally, we apply RHIG to a well-established stochastic prediction error model and provide expected regret and concentration bounds under correlated prediction errors. Lastly, we numerically test the performance of RHIG on quadrotor tracking problems.

[Learning the Linear Quadratic Regulator from Nonlinear Observations](#)

- Zakaria Mhammedi, Dylan J. Foster, Max Simchowitz, Dipendra Misra, Wen Sun, Akshay Krishnamurthy, Alexander Rakhlin, John Langford
- abstract@[open-review](#): We introduce a new problem setting for continuous control called the LQR with Rich Observations, or RichLQR. In our setting, the environment is summarized by a low-dimensional continuous latent state with linear dynamics and quadratic costs, but the agent operates on high-dimensional, nonlinear observations such as images from a camera. To enable sample-efficient learning, we assume that the learner has access to a class of decoder functions (e.g., neural networks) that is flexible enough to capture the mapping from observations to latent states. We introduce a new algorithm, RichID, which learns a near-optimal policy for the RichLQR with sample complexity scaling only with the dimension of the latent state space and the capacity of the decoder function class. RichID is oracle-efficient and accesses the decoder class only through calls to a least-squares regression oracle. To

our knowledge, our results constitute the first provable sample complexity guarantee for continuous control with an unknown nonlinearity in the system model.

[Reconciling Modern Deep Learning with Traditional Optimization Analyses: The Intrinsic Learning Rate](#)

- Zhiyuan Li, Kaifeng Lyu, Sanjeev Arora
- abstract@[open-review](#): Recent works (e.g., (Li \& Arora, 2020)) suggest that the use of popular normalization schemes (including Batch Normalization) in today's deep learning can move it far from a traditional optimization viewpoint, e.g., use of exponentially increasing learning rates. The current paper highlights other ways in which behavior of normalized nets departs from traditional viewpoints, and then initiates a formal framework for studying their mathematics via suitable adaptation of the conventional framework namely, modeling SGD-induced training trajectory via a suitable stochastic differential equation (SDE) with a noise term that captures gradient noise. This yields: (a) A new \text{intrinsic learning rate} parameter that is the product of the normal learning rate η and weight decay factor λ . Analysis of the SDE shows how the effective speed of learning varies and equilibrates over time under the control of intrinsic LR. (b) A challenge---via theory and experiments---to popular belief that good generalization requires large learning rates at the start of training. (c) New experiments, backed by mathematical intuition, suggesting the number of steps to equilibrium (in function space) scales as the inverse of the intrinsic learning rate, as opposed to the exponential time convergence bound implied by SDE analysis. We name it the \text{Fast Equilibrium Conjecture} and suggest it holds the key to why Batch Normalization is effective.

[Scalable Graph Neural Networks via Bidirectional Propagation](#)

- Ming Chen, Zhewei Wei, Bolin Ding, Yaliang Li, Ye Yuan, Xiaoyong Du, Ji-Rong Wen
- abstract@[open-review](#): Graph Neural Networks (GNN) are an emerging field for learning on non-Euclidean data. Recently, there has been increased interest in designing GNN that scales to large graphs. Most existing methods use "graph sampling" or "layer-wise sampling" techniques to reduce training time; However, these methods still suffer from degrading performance and scalability problems when applying to graphs with billions of edges. In this paper, we present GBP, a scalable GNN that utilizes a localized bidirectional propagation process from both the feature vector and the training/testing nodes. Theoretical analysis shows that GBP is the first method that achieves sub-linear time complexity for both the precomputation and the training phases. An extensive empirical study demonstrates that GBP achieves state-of-the-art performance with significantly less training/testing time. Most notably, GBP is able to deliver superior performance on a graph with over 60 million nodes and 1.8 billion edges in less than 2,000 seconds on a single machine.

[Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning](#)

- Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, Jinwoo Shin
- abstract@[open-review](#): While semi-supervised learning (SSL) has proven to be a promising way for leveraging unlabeled data when labeled data is scarce, the existing SSL algorithms typically assume that training class distributions are balanced. However, these SSL algorithms trained under imbalanced class distributions can severely suffer when generalizing to a balanced testing criterion, since they utilize biased pseudo-labels of unlabeled data toward majority classes. To alleviate this issue, we formulate a convex optimization problem to softly refine the pseudo-labels generated from the biased model, and develop a simple algorithm, named Distribution Aligning Refinery of Pseudo-label (DARP) that solves it provably and efficiently. Under various class imbalanced semi-supervised scenarios, we demonstrate the effectiveness of DARP and its compatibility with state-of-the-art SSL schemes.

[Assisted Learning: A Framework for Multi-Organization Learning](#)

- Xun Xian, Xinran Wang, Jie Ding, Reza Ghanadan
- abstract@[open-review](#): In an increasing number of AI scenarios, collaborations among different organizations or agents (e.g., human and robots, mobile units) are often essential to accomplish an organization-specific mission. However, to avoid leaking useful and possibly proprietary information, organizations typically enforce stringent security constraints on sharing modeling algorithms and data, which significantly limits collaborations. In this work, we introduce the Assisted Learning framework for organizations to assist each other in supervised learning tasks without revealing any organization's algorithm, data, or even task. An organization seeks assistance by broadcasting task-specific but nonsensitive statistics and incorporating others' feedback in one or more iterations to eventually improve its predictive performance. Theoretical and experimental studies, including real-world medical benchmarks, show that Assisted Learning can often achieve near-oracle learning performance as if data and training processes were centralized.

[The Strong Screening Rule for SLOPE](#)

- Johan Larsson, Małgorzata Bogdan, Jonas Wallin
- abstract@[open-review](#): Extracting relevant features from data sets where the number of observations (n) is much smaller than the number of predictors (p) is a major challenge in modern statistics. Sorted L-One Penalized Estimation (SLOPE)---a generalization of the lasso---is a promising method within this setting. Current numerical procedures for SLOPE, however, lack the efficiency that respective tools for the lasso enjoy, particularly in the context of estimating a complete regularization path. A key component in the efficiency of the lasso is predictor screening rules: rules that allow predictors to be discarded before estimating the model. This is the first paper to establish such a rule for SLOPE. We develop a screening rule for SLOPE by examining its subdifferential and show that this rule is a generalization of the strong rule for the lasso. Our rule is heuristic, which means that it may discard predictors erroneously. In our paper, however, we show that such situations are rare and easily safeguarded against by a simple check of the optimality conditions. Our numerical experiments show that the rule performs well in practice, leading to improvements by orders of magnitude for data in the ($p \gg n$) domain, as well as incurring no additional computational overhead when ($n > p$).

[STLnet: Signal Temporal Logic Enforced Multivariate Recurrent Neural Networks](#)

- Meiyi Ma, Ji Gao, Lu Feng, John Stankovic
- abstract@[open-review](#): Recurrent Neural Networks (RNNs) have made great achievements for sequential prediction tasks. In practice, the target sequence often follows certain model properties or patterns (e.g., reasonable ranges, consecutive changes, resource constraint, temporal correlations between multiple variables, existence, unusual cases, etc.). However, RNNs cannot guarantee their learned distributions satisfy these model properties. It is even more challenging for predicting large-scale and complex Cyber-Physical Systems. Failure to produce outcomes that meet these model properties will result in inaccurate and even meaningless results. In this paper, we develop a new temporal logic-based learning framework, STLnet, which guides the RNN learning process with auxiliary knowledge of model properties, and produces a more robust model for improved future predictions. Our framework can be applied to general sequential deep learning models, and trained in an end-to-end manner with back-propagation. We evaluate the performance of STLnet using large-scale real-world city data. The experimental results show STLnet not only improves the accuracy of predictions, but importantly also guarantees the satisfaction of model properties and increases the robustness of RNNs.

[Election Coding for Distributed Learning: Protecting SignSGD against Byzantine Attacks](#)

- Jy-yong Sohn, Dong-Jun Han, Beongjun Choi, Jaekyun Moon
- abstract@[open-review](#): Current distributed learning systems suffer from serious performance degradation under Byzantine attacks. This paper proposes Election Coding, a coding-theoretic framework to guarantee Byzantine-robustness for distributed learning algorithms based on signed stochastic gradient

descent (SignSGD) that minimizes the worker-master communication load. The suggested framework explores new information-theoretic limits of finding the majority opinion when some workers could be attacked by adversary, and paves the road to implement robust and communication-efficient distributed learning algorithms. Under this framework, we construct two types of codes, random Bernoulli codes and deterministic algebraic codes, that tolerate Byzantine attacks with a controlled amount of computational redundancy and guarantee convergence in general non-convex scenarios. For the Bernoulli codes, we provide an upper bound on the error probability in estimating the signs of the true gradients, which gives useful insights into code design for Byzantine tolerance. The proposed deterministic codes are proven to perfectly tolerate arbitrary Byzantine attacks. Experiments on real datasets confirm that the suggested codes provide substantial improvement in Byzantine tolerance of distributed learning systems employing SignSGD.

[Reducing Adversarially Robust Learning to Non-Robust PAC Learning](#)

- Omar Montasser, Steve Hanneke, Nati Srebro
- abstract@[open-review](#): We study the problem of reducing adversarially robust learning to standard PAC learning, i.e. the complexity of learning adversarially robust predictors using access to only a black-box non-robust learner. We give a reduction that can robustly learn any hypothesis class C using any non-robust learner A for C . The number of calls to A depends logarithmically on the number of allowed adversarial perturbations per example, and we give a lower bound showing this is unavoidable.

[Top-k Training of GANs: Improving GAN Performance by Throwing Away Bad Samples](#)

- Samarth Sinha, Zhengli Zhao, Anirudh Goyal ALIAS PARTH GOYAL, Colin A. Raffel, Augustus Odena
- abstract@[open-review](#): We introduce a simple (one line of code) modification to the Generative Adversarial Network (GAN) training algorithm that materially improves results with no increase in computational cost. When updating the generator parameters, we simply zero out the gradient contributions from the elements of the batch that the critic scores as least realistic'. Through experiments on many different GAN variants, we show that this 'top-k update' procedure is a generally applicable improvement. In order to understand the nature of the improvement, we conduct extensive analysis on a simple mixture-of-Gaussians dataset and discover several interesting phenomena. Among these is that, when gradient updates are computed using the worst-scoring batch elements, samples can actually be pushed further away from their nearest mode. We also apply our method to state-of-the-art GAN models including BigGAN and improve state-of-the-art FID for conditional generation on CIFAR-10 from 9.21 to 8.57.

[Black-Box Optimization with Local Generative Surrogates](#)

- Sergey Shirobokov, Vladislav Belavin, Michael Kagan, Andrei Ustyuzhanin, Atilim Gunes Baydin
- abstract@[open-review](#): We propose a novel method for gradient-based optimization of black-box simulators using differentiable local surrogate models. In fields such as physics and engineering, many processes are modeled with non-differentiable simulators with intractable likelihoods. Optimization of these forward models is particularly challenging, especially when the simulator is stochastic. To address such cases, we introduce the use of deep generative models to iteratively approximate the simulator in local neighborhoods of the parameter space. We demonstrate that these local surrogates can be used to approximate the gradient of the simulator, and thus enable gradient-based optimization of simulator parameters. In cases where the dependence of the simulator on the parameter space is constrained to a low dimensional submanifold, we observe that our method attains minima faster than baseline methods, including Bayesian optimization, numerical optimization and approaches using score function gradient estimators.

[Efficient Generation of Structured Objects with Constrained Adversarial Networks](#)

- Luca Di Liello, Pierfrancesco Ardino, Jacopo Gobbi, Paolo Morettin, Stefano Teso, Andrea Passerini
- abstract@[open-review](#): Generative Adversarial Networks (GANs) struggle to generate structured objects like molecules and game maps. The issue is that structured objects must satisfy hard requirements (e.g., molecules must be chemically valid) that are difficult to acquire from examples alone. As a remedy, we propose Constrained Adversarial Networks (CANs), an extension of GANs in which the constraints are embedded into the model during training. This is achieved by penalizing the generator proportionally to the mass it allocates to invalid structures. In contrast to other generative models, CANs support efficient inference of valid structures (with high probability) and allows to turn on and off the learned constraints at inference time. CANs handle arbitrary logical constraints and leverage knowledge compilation techniques to efficiently evaluate the disagreement between the model and the constraints. Our setup is further extended to hybrid logical-neural constraints for capturing very complex constraints, like graph reachability. An extensive empirical analysis shows that CANs efficiently generate valid structures that are both high-quality and novel.

[Hard Example Generation by Texture Synthesis for Cross-domain Shape Similarity Learning](#)

- Huan Fu, Shunming Li, Rongfei Jia, Mingming Gong, Binqiang Zhao, Dacheng Tao
- abstract@[open-review](#): Image-based 3D shape retrieval (IBSR) aims to find the corresponding 3D shape of a given 2D image from a large 3D shape database. The common routine is to map 2D images and 3D shapes into an embedding space and define (or learn) a shape similarity measure. While metric learning with some adaptation techniques seems to be a natural solution to shape similarity learning, the performance is often unsatisfactory for fine-grained shape retrieval. In the paper, we identify the source of the poor performance and propose a practical solution to this problem. We find that the shape difference between a negative pair is entangled with the texture gap, making metric learning ineffective in pushing away negative pairs. To tackle this issue, we develop a geometry-focused multi-view metric learning framework empowered by texture synthesis. The synthesis of textures for 3D shape models creates hard triplets, which suppress the adverse effects of rich texture in 2D images, thereby push the network to focus more on discovering geometric characteristics. Our approach shows state-of-the-art performance on a recently released large-scale 3D-FUTURE [1] repository, as well as three widely studied benchmarks, including Pix3D [2], Stanford Cars [3], and Comp Cars [4]. Codes will be made publicly available at: <https://github.com/3D-FRONT-FUTURE/IBSR-texture>.

[Recovery of sparse linear classifiers from mixture of responses](#)

- Venkata Gandikota, Arya Mazumdar, Soumyabrata Pal
- abstract@[open-review](#): In the problem of learning a mixture of linear classifiers, the aim is to learn a collection of hyperplanes from a sequence of binary responses. Each response is a result of querying with a vector and indicates the side of a randomly chosen hyperplane from the collection the query vector belongs to. This model is quite rich while dealing with heterogeneous data with categorical labels and has only been studied in some special settings. We look at a hitherto unstudied problem of query complexity upper bound of recovering all the hyperplanes, especially for the case when the hyperplanes are sparse. This setting is a natural generalization of the extreme quantization problem known as 1-bit compressed sensing. Suppose we have a set of l unknown k -sparse vectors. We can query the set with another vector a , to obtain the sign of the inner product of a and a randomly chosen vector from the l -set. How many queries are sufficient to identify all the l unknown vectors? This question is significantly more challenging than both the basic 1-bit compressed sensing problem (i.e., $l = 1$ case) and the analogous regression problem (where the value instead of the sign is provided). We provide rigorous query complexity results (with efficient algorithms) for this problem.

[Efficient Distance Approximation for Structured High-Dimensional Distributions via Learning](#)

- Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S Meel, N. V. Vinodchandran

- abstract@[open-review](#): We design efficient distance approximation algorithms for several classes of well-studied structured high-dimensional distributions. Specifically, we present algorithms for the following problems (where dTV is the total variation distance):

Given sample access to two Bayesian networks P1 and P2 over known directed acyclic graphs G1 and G2 having n nodes and bounded in-degree, approximate dTV(P1, P2) to within additive error $\hat{\mu}$ using $\text{poly}(n, 1/\hat{\mu})$ samples and time. Given sample access to two ferromagnetic Ising models P1 and P2 on n variables with bounded width, approximate dTV(P1, P2) to within additive error $\hat{\mu}$ using $\text{poly}(n, 1/\hat{\mu})$ samples and time. Given sample access to two n-dimensional Gaussians P1 and P2, approximate dTV(P1, P2) to within additive error $\hat{\mu}$ using $\text{poly}(n, 1/\hat{\mu})$ samples and time. Given access to observations from two causal models P and Q on n variables that are defined over known causal graphs, approximate dTV(Pa, Qa) to within additive error $\hat{\mu}$ using $\text{poly}(n, 1/\hat{\mu})$ samples and time, where Pa and Qa are the interventional distributions obtained by the intervention $\text{do}(A = a)$ on P and Q respectively for a particular variable A.

The distance approximation algorithms immediately imply new tolerant closeness testers for the corresponding classes of distributions. Prior to our work, only non-tolerant testers were known for both Bayes net distributions and Ising models, and no testers with quantitative guarantee were known for interventional distributions. To best of our knowledge, efficient distance approximation algorithms for Gaussian distributions were not present in the literature. Our algorithms are designed using a conceptually simple but general framework that is applicable to a variety of scenarios.

[A Single Recipe for Online Submodular Maximization with Adversarial or Stochastic Constraints](#)

- Omid Sadeghi, Prasanna Raut, Maryam Fazel
- abstract@[open-review](#): In this paper, we consider an online optimization problem in which the reward functions are DR-submodular, and in addition to maximizing the total reward, the sequence of decisions must satisfy some convex constraints on average. Specifically, at each round $t \in \{1, \dots, T\}$, upon committing to an action x_t , a DR-submodular utility function $f_t(\cdot)$ and a convex constraint function $g_t(\cdot)$ are revealed, and the goal is to maximize the overall utility while ensuring the average of the constraint functions $\frac{1}{T} \sum_{t=1}^T g_t(x_t)$ is non-positive. Such cumulative constraints arise naturally in applications where the average resource consumption is required to remain below a prespecified threshold. We study this problem under an adversarial model and a stochastic model for the convex constraints, where the functions g_t can vary arbitrarily or according to an i.i.d. process over time slots $\{1, \dots, T\}$, respectively. We propose a single algorithm which achieves sub-linear (with respect to T) regret as well as sub-linear constraint violation bounds in both settings, without prior knowledge of the regime. Prior works have studied this problem in the special case of linear constraint functions. Our results not only improve upon the existing bounds under linear cumulative constraints, but also give the first sub-linear bounds for general convex long-term constraints.

[Learning Sparse Prototypes for Text Generation](#)

- Junxian He, Taylor Berg-Kirkpatrick, Graham Neubig
- abstract@[open-review](#): Prototype-driven text generation uses non-parametric models that first choose from a library of sentence "prototypes" and then modify the prototype to generate the output text. While effective, these methods are inefficient at test time as a result of needing to store and index the entire training corpus. Further, existing methods often require heuristics to identify which prototypes to reference at training time. In this paper, we propose a novel generative model that automatically learns a sparse prototype support set that, nonetheless, achieves strong language modeling performance. This is achieved by (1) imposing a sparsity-inducing prior on the prototype selection distribution, and (2) utilizing amortized variational inference to learn a prototype retrieval function. In experiments, our model outperforms previous prototype-driven language models while achieving up to a 1000x memory reduction, as well as a 1000x speed-up at test time. More interestingly, we show that the learned prototypes are able to capture semantics and syntax at different granularity as we vary the sparsity of prototype selection, and that certain sentence attributes can be controlled by specifying the prototype for generation.

[Implicit Rank-Minimizing Autoencoder](#)

- Li Jing, Jure Zbontar, yann lecun
- abstract@[open-review](#): An important component of autoencoder methods is the method by which the information capacity of the latent representation is minimized or limited. In this work, the rank of the covariance matrix of the codes is implicitly minimized by relying on the fact that gradient descent learning in multi-layer linear networks leads to minimum-rank solutions. By inserting a number of extra linear layers between the encoder and the decoder, the system spontaneously learns representations with a low effective dimension. The model, dubbed Implicit Rank-Minimizing Autoencoder (IRMAE), is simple, deterministic, and learns continuous latent space. We demonstrate the validity of the method on several image generation and representation learning tasks.

[Storage Efficient and Dynamic Flexible Runtime Channel Pruning via Deep Reinforcement Learning](#)

- Jianda Chen, Shangyu Chen, Sinno Jialin Pan
- abstract@[open-review](#): In this paper, we propose a deep reinforcement learning (DRL) based framework to efficiently perform runtime channel pruning on convolutional neural networks (CNNs). Our DRL-based framework aims to learn a pruning strategy to determine how many and which channels to be pruned in each convolutional layer, depending on each individual input instance at runtime. Unlike existing runtime pruning methods which require to store all channels parameters for inference, our framework can reduce parameters storage consumption by introducing a static pruning component. Comparison experimental results with existing runtime and static pruning methods on state-of-the-art CNNs demonstrate that our proposed framework is able to provide a tradeoff between dynamic flexibility and storage efficiency in runtime channel pruning.

[Task-Oriented Feature Distillation](#)

- Linfeng Zhang, Yukang Shi, Zuoqiang Shi, Kaisheng Ma, Chenglong Bao
- abstract@[open-review](#): Feature distillation, a primary method in knowledge distillation, always leads to significant accuracy improvements. Most existing methods distill features in the teacher network through a manually designed transformation. In this paper, we propose a novel distillation method named task-oriented feature distillation (TOFD) where the transformation is convolutional layers that are trained in a data-driven manner by task loss. As a result, the task-oriented information in the features can be captured and distilled to students. Moreover, an orthogonal loss is applied to the feature resizing layer in TOFD to improve the performance of knowledge distillation. Experiments show that TOFD outperforms other distillation methods by a large margin on both image classification and 3D classification tasks. Codes have been released in Github.

[Entropic Causal Inference: Identifiability and Finite Sample Results](#)

- Spencer Compton, Murat Kocaoglu, Kristjan Greenewald, Dmitriy Katz
- abstract@[open-review](#): Entropic causal inference is a framework for inferring the causal direction between two categorical variables from observational data. The central assumption is that the amount of unobserved randomness in the system is not too large. This unobserved randomness is measured by the entropy of the exogenous variable in the underlying structural causal model, which governs the causal relation between the observed variables. Kocaoglu et al. conjectured that the causal direction is identifiable when the entropy of the exogenous variable is not too large. In this paper, we prove a variant of their conjecture. Namely, we show that for almost all causal models where the exogenous variable has entropy that does not scale with the number of states of the observed variables, the causal direction is identifiable from observational data. We also consider the minimum entropy coupling-based algorithmic approach presented by Kocaoglu et al., and for the first time demonstrate algorithmic identifiability guarantees using a finite number of

samples. We conduct extensive experiments to evaluate the robustness of the method to relaxing some of the assumptions in our theory and demonstrate that both the constant-entropy exogenous variable and the no latent confounder assumptions can be relaxed in practice. We also empirically characterize the number of observational samples needed for causal identification. Finally, we apply the algorithm on Tuebingen cause-effect pairs dataset.

[Rewriting History with Inverse RL: Hindsight Inference for Policy Improvement](#)

- Ben Eysenbach, XINYANG GENG, Sergey Levine, Russ R. Salakhutdinov
- abstract@[open-review](#): Multi-task reinforcement learning (RL) aims to simultaneously learn policies for solving many tasks. Several prior works have found that relabeling past experience with different reward functions can improve sample efficiency. Relabeling methods typically pose the question: if, in hindsight, we assume that our experience was optimal for some task, for what task was it optimal? Inverse RL answers this question. In this paper we show that inverse RL is a principled mechanism for reusing experience across tasks. We use this idea to generalize goal-relabeling techniques from prior work to arbitrary types of reward functions. Our experiments confirm that relabeling data using inverse RL outperforms prior relabeling methods on goal-reaching tasks, and accelerates learning on more general multi-task settings where prior methods are not applicable, such as domains with discrete sets of rewards and those with linear reward functions.

[Variance-Reduced Off-Policy TDC Learning: Non-Asymptotic Convergence Analysis](#)

- Shaocong Ma, Yi Zhou, Shaofeng Zou
- abstract@[open-review](#): Variance reduction techniques have been successfully applied to temporal-difference (TD) learning and help to improve the sample complexity in policy evaluation. However, the existing work applied variance reduction to either the less popular one time-scale TD algorithm or the two time-scale GTD algorithm but with a finite number of i.i.d.\ samples, and both algorithms apply to only the on-policy setting. In this work, we develop a variance reduction scheme for the two time-scale TDC algorithm in the off-policy setting and analyze its non-asymptotic convergence rate over both i.i.d.\ and Markovian samples. In the i.i.d setting, our algorithm achieves an improved sample complexity $\mathcal{O}(\epsilon^{-\frac{3}{5}} \log \frac{1}{\epsilon})$ over the state-of-the-art result $\mathcal{O}(\epsilon^{-1} \log \frac{1}{\epsilon})$. In the Markovian setting, our algorithm achieves the state-of-the-art sample complexity $\mathcal{O}(\epsilon^{-1} \log \frac{1}{\epsilon})$ that is near-optimal. Experiments demonstrate that the proposed variance-reduced TDC achieves a smaller asymptotic convergence error than both the conventional TDC and the variance-reduced TD.

[AdaTune: Adaptive Tensor Program Compilation Made Efficient](#)

- Menghao Li, Minjia Zhang, Chi Wang, Mingqin Li
- abstract@[open-review](#): Deep learning models are computationally intense, and implementations often have to be highly optimized by experts or hardware vendors to be usable in practice. The DL compiler, together with Learning to Compile have proven to be a powerful technique for optimizing tensor programs. However, a limitation of this approach is that it still suffers from unbearably long overall optimization time. In this paper, we present a new method, called AdaTune, that significantly reduces the optimization time of tensor programs for high-performance deep learning inference. In particular, we propose an adaptive evaluation method that statistically early terminates a costly hardware measurement without losing much accuracy. We further devise a surrogate model with uncertainty quantification that allows the optimization to adapt to hardware and model heterogeneity better. Finally, we introduce a contextual optimizer that provides adaptive control of the exploration and exploitation to improve the transformation space searching effectiveness. We evaluate and compare the levels of optimization obtained by a state-of-the-art DL compiler and AdaTune. The experiment results show that AdaTune obtains up to 115% higher GFLOPS than the baseline under the same optimization time budget. Furthermore, AdaTune provides 1.3--3.9X speedup in optimization time over the state-of-the-art to reach the same optimization quality for a range of models across different hardware architectures.

[When Do Neural Networks Outperform Kernel Methods?](#)

- Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, Andrea Montanari
- abstract@[open-review](#): For a certain scaling of the initialization of stochastic gradient descent (SGD), wide neural networks (NN) have been shown to be well approximated by reproducing kernel Hilbert space (RKHS) methods. Recent empirical work showed that, for some classification tasks, RKHS methods can replace NNs without a large loss in performance. On the other hand, two-layers NNs are known to encode richer smoothness classes than RKHS and we know of special examples for which SGD-trained NN provably outperform RKHS. This is true even in the wide network limit, for a different scaling of the initialization. How can we reconcile the above claims? For which tasks do NNs outperform RKHS? If covariates are nearly isotropic, RKHS methods suffer from the curse of dimensionality, while NNs can overcome it by learning the best low-dimensional representation. Here we show that this curse of dimensionality becomes milder if the covariates display the same low-dimensional structure as the target function, and we precisely characterize this tradeoff. Building on these results, we present the spiked covariates model that can capture in a unified framework both behaviors observed in earlier works. We hypothesize that such a latent low-dimensional structure is present in image classification. We numerically test this hypothesis by showing that specific perturbations of the training distribution degrade the performances of RKHS methods much more significantly than NNs.

[STEER : Simple Temporal Regularization For Neural ODE](#)

- Arnab Ghosh, Harkirat Behl, Emilien Dupont, Philip Torr, Vinay Namboodiri
- abstract@[open-review](#): Training Neural Ordinary Differential Equations (ODEs) is often computationally expensive. Indeed, computing the forward pass of such models involves solving an ODE which can become arbitrarily complex during training. Recent works have shown that regularizing the dynamics of the ODE can partially alleviate this. In this paper we propose a new regularization technique: randomly sampling the end time of the ODE during training. The proposed regularization is simple to implement, has negligible overhead and is effective across a wide variety of tasks. Further, the technique is orthogonal to several other methods proposed to regularize the dynamics of ODEs and as such can be used in conjunction with them. We show through experiments on normalizing flows, time series models and image recognition that the proposed regularization can significantly decrease training time and even improve performance over baseline models.

[A Variational Approach for Learning from Positive and Unlabeled Data](#)

- Hui Chen, Fangqing Liu, Yin Wang, Liyue Zhao, Hao Wu
- abstract@[open-review](#): Learning binary classifiers only from positive and unlabeled (PU) data is an important and challenging task in many real-world applications, including web text classification, disease gene identification and fraud detection, where negative samples are difficult to verify experimentally. Most recent PU learning methods are developed based on the misclassification risk of the supervised learning type, and they may suffer from inaccurate estimates of class prior probabilities. In this paper, we introduce a variational principle for PU learning that allows us to quantitatively evaluate the modeling error of the Bayesian classifier directly from given data. This leads to a loss function which can be efficiently calculated without involving class prior estimation or any other intermediate estimation problems, and the variational learning method can then be employed to optimize the classifier under general conditions. We illustrate the effectiveness of the proposed variational method on a number of benchmark examples.

[Efficient Clustering Based On A Unified View Of K-means And Ratio-cut](#)

- Shenfei Pei, Feiping Nie, Rong Wang, Xuelong Li
- abstract@[open-review](#): Spectral clustering and \$k\$-means, both as two major traditional clustering methods, are still attracting a lot of attention, although a variety of novel clustering algorithms have been proposed in recent years. Firstly, a unified framework of \$k\$-means and ratio-cut is revisited, and a novel and efficient clustering algorithm is then proposed based on this framework. The time and space complexity of our method are both linear with respect to the number of samples, and are independent of the number of clusters to construct, more importantly. These properties mean that it is easily scalable and applicable to large practical problems. Extensive experiments on 12 real-world benchmark and 8 facial datasets validate the advantages of the proposed algorithms compared to the state-of-the-art clustering algorithms. In particular, over 15x and 7x speed-up can be obtained with respect to \$k\$-means on the synthetic dataset of 1 million samples and the benchmark dataset (CelebA) of 200k samples, respectively [GitHub].

Recurrent Switching Dynamical Systems Models for Multiple Interacting Neural Populations

- Joshua Glaser, Matthew Whiteway, John P. Cunningham, Liam Paninski, Scott Linderman
- abstract@[open-review](#): Modern recording techniques can generate large-scale measurements of multiple neural populations over extended time periods. However, it remains a challenge to model non-stationary interactions between high-dimensional populations of neurons. To tackle this challenge, we develop recurrent switching linear dynamical systems models for multiple populations. Here, each high-dimensional neural population is represented by a unique set of latent variables, which evolve dynamically in time. Populations interact with each other through this low-dimensional space. We allow the nature of these interactions to change over time by using a discrete set of dynamical states. Additionally, we parameterize these discrete state transition rules to capture which neural populations are responsible for switching between interaction states. To fit the model, we use variational expectation-maximization with a structured mean-field approximation. After validating the model on simulations, we apply it to two different neural datasets: spiking activity from motor areas in a non-human primate, and calcium imaging from neurons in the nematode *C. elegans*. In both datasets, the model reveals behaviorally-relevant discrete states with unique inter-population interactions and different populations that predict transitioning between these states.

Coresets via Bilevel Optimization for Continual Learning and Streaming

- Zalán Borsos, Mojmir Mutny, Andreas Krause
- abstract@[open-review](#): Coresets are small data summaries that are sufficient for model training. They can be maintained online, enabling efficient handling of large data streams under resource constraints. However, existing constructions are limited to simple models such as k-means and logistic regression. In this work, we propose a novel coreset construction via cardinality-constrained bilevel optimization. We show how our framework can efficiently generate coresets for deep neural networks, and demonstrate its empirical benefits in continual learning and in streaming settings.

Generalized Independent Noise Condition for Estimating Latent Variable Causal Graphs

- Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, Kun Zhang
- abstract@[open-review](#): Causal discovery aims to recover causal structures or models underlying the observed data. Despite its success in certain domains, most existing methods focus on causal relations between observed variables, while in many scenarios the observed ones may not be the underlying causal variables (e.g., image pixels), but are generated by latent causal variables or confounders that are causally related. To this end, in this paper, we consider Linear, Non-Gaussian Latent variable Models (LiNGLaMs), in which latent confounders are also causally related, and propose a Generalized Independent Noise (GIN) condition to estimate such latent variable graphs. Specifically, for two observed random vectors \mathbf{Y} and \mathbf{Z} , GIN holds if and only if $\Omega^{\text{intercausal}}(\mathbf{Y})$ and \mathbf{Z} are statistically independent, where Ω is a parameter vector characterized from the cross-covariance between \mathbf{Y} and \mathbf{Z} . From the graphical view, roughly speaking, GIN implies that causally earlier latent common causes of variables in \mathbf{Y} d-separate \mathbf{Y} from \mathbf{Z} . Interestingly, we find that the independent noise condition, i.e., if there is no confounder, causes are independent from the error of regressing the effect on the causes, can be seen as a special case of GIN. Moreover, we show that GIN helps locate latent variables and identify their causal structure, including causal directions. We further develop a recursive learning algorithm to achieve these goals. Experimental results on synthetic and real-world data demonstrate the effectiveness of our method.

Understanding and Exploring the Network with Stochastic Architectures

- Zhijie Deng, Yinpeng Dong, Shifeng Zhang, Jun Zhu
- abstract@[open-review](#): There is an emerging trend to train a network with stochastic architectures to enable various architectures to be plugged and played during inference. However, the existing investigation is highly entangled with neural architecture search (NAS), limiting its widespread use across scenarios. In this work, we decouple the training of a network with stochastic architectures (NSA) from NAS and provide a first systematical investigation on it as a stand-alone problem. We first uncover the characteristics of NSA in various aspects ranging from training stability, convergence, predictive behaviour, to generalization capacity to unseen architectures. We identify various issues of the vanilla NSA, such as training/test disparity and function mode collapse, and further propose the solutions to these issues with theoretical and empirical insights. We believe that these results could also serve as good heuristics for NAS. Given these understandings, we further apply the NSA with our improvements into diverse scenarios to fully exploit its promise of inference-time architecture stochasticity, including model ensemble, uncertainty estimation and semi-supervised learning. Remarkable performance (e.g., 2.75% error rate and 0.0032 expected calibration error on CIFAR-10) validate the effectiveness of such a model, providing new perspectives of exploring the potential of the network with stochastic architectures, beyond NAS.

All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation

- jean barbier, Nicolas Macris, Cynthia Rush
- abstract@[open-review](#): We determine statistical and computational limits for estimation of a rank-one matrix (the spike) corrupted by an additive gaussian noise matrix, in a sparse limit, where the underlying hidden vector (that constructs the rank-one matrix) has a number of non-zero components that scales sub-linearly with the total dimension of the vector, and the signal-to-noise ratio tends to infinity at an appropriate speed. We prove explicit low-dimensional variational formulas for the asymptotic mutual information between the spike and the observed noisy matrix and analyze the approximate message passing algorithm in the sparse regime. For Bernoulli and Bernoulli-Rademacher distributed vectors, and when the sparsity and signal strength satisfy an appropriate scaling relation, we find all-or-nothing phase transitions for the asymptotic minimum and algorithmic mean-square-errors. These jump from their maximum possible value to zero, at well defined signal-to-noise thresholds whose asymptotic values we determine exactly. In the asymptotic regime the statistical-to-algorithmic gap diverges indicating that sparse recovery is hard for approximate message passing.

Deep Evidential Regression

- Alexander Amini, Wilko Schwarting, Ava Soleimany, Daniela Rus
- abstract@[open-review](#): Deterministic neural networks (NNs) are increasingly being deployed in safety critical domains, where calibrated, robust, and efficient measures of uncertainty are crucial. In this paper, we propose a novel method for training non-Bayesian NNs to estimate a continuous target as well as its associated evidence in order to learn both aleatoric and epistemic uncertainty. We accomplish this by placing evidential priors over the original Gaussian likelihood function and training the NN to infer the hyperparameters of the evidential distribution. We additionally impose priors during training such that the model is regularized when its predicted evidence is not aligned with the correct output. Our method does not rely on sampling during inference or on out-of-distribution (OOD) examples for training, thus enabling efficient and scalable uncertainty learning. We demonstrate learning well-

calibrated measures of uncertainty on various benchmarks, scaling to complex computer vision tasks, as well as robustness to adversarial and OOD test samples.

[Analytical Probability Distributions and Exact Expectation-Maximization for Deep Generative Networks](#)

- Randall Balestiero, Sebastien PARIS, Richard Baraniuk
- abstract@[open-review](#): Deep Generative Networks (DGNs) with probabilistic modeling of their output and latent space are currently trained via Variational Autoencoders (VAEs). In the absence of a known analytical form for the posterior and likelihood expectation, VAEs resort to approximations, including (Amortized) Variational Inference (AVI) and Monte-Carlo sampling. We exploit the Continuous Piecewise Affine property of modern DGNs to derive their posterior and marginal distributions as well as the latter's first two moments. These findings enable us to derive an analytical Expectation-Maximization (EM) algorithm for gradient-free DGN learning. We demonstrate empirically that EM training of DGNs produces greater likelihood than VAE training. Our new framework will guide the design of new VAE AVI that better approximates the true posterior and open new avenues to apply standard statistical tools for model comparison, anomaly detection, and missing data imputation.

[Bayesian Pseudocoresets](#)

- Dionysis Manousakas, Zuheng Xu, Cecilia Mascolo, Trevor Campbell
- abstract@[open-review](#): Standard Bayesian inference algorithms are prohibitively expensive in the regime of modern large-scale data. Recent work has found that a small, weighted subset of data (a coresnet) may be used in place of the full dataset during inference, taking advantage of data redundancy to reduce computational cost. However, this approach has limitations in the increasingly common setting of sensitive, high-dimensional data. Indeed, we prove that there are situations in which the Kullback-Leibler divergence between the optimal coresnet and the true posterior grows with data dimension; and as coresnets include a subset of the original data, they cannot be constructed in a manner that preserves individual privacy. We address both of these issues with a single unified solution, Bayesian pseudocoresets --- a small weighted collection of synthetic "pseudodata"---along with a variational optimization method to select both pseudodata and weights. The use of pseudodata (as opposed to the original datapoints) enables both the summarization of high-dimensional data and the differentially private summarization of sensitive data. Real and synthetic experiments on high-dimensional data demonstrate that Bayesian pseudocoresets achieve significant improvements in posterior approximation error compared to traditional coresets, and that pseudocoresets provide privacy without a significant loss in approximation quality.

[See, Hear, Explore: Curiosity via Audio-Visual Association](#)

- Victoria Dean, Shubham Tulsiani, Abhinav Gupta
- abstract@[open-review](#): Exploration is one of the core challenges in reinforcement learning. A common formulation of curiosity-driven exploration uses the difference between the real future and the future predicted by a learned model. However, predicting the future is an inherently difficult task which can be ill-posed in the face of stochasticity. In this paper, we introduce an alternative form of curiosity that rewards novel associations between different senses. Our approach exploits multiple modalities to provide a stronger signal for more efficient exploration. Our method is inspired by the fact that, for humans, both sight and sound play a critical role in exploration. We present results on several Atari environments and Habitat (a photorealistic navigation simulator), showing the benefits of using an audio-visual association model for intrinsically guiding learning agents in the absence of external rewards. For videos and code, see <https://vdean.github.io/audio-curiosity.html>.

[Adversarial Training is a Form of Data-dependent Operator Norm Regularization](#)

- Kevin Roth, Yannic Kilcher, Thomas Hofmann
- abstract@[open-review](#): We establish a theoretical link between adversarial training and operator norm regularization for deep neural networks. Specifically, we prove that $\| \cdot \|_p$ -norm constrained projected gradient ascent based adversarial training with an $\| \cdot \|_q$ -norm loss on the logits of clean and perturbed inputs is equivalent to data-dependent (p, q) operator norm regularization. This fundamental connection confirms the long-standing argument that a network's sensitivity to adversarial examples is tied to its spectral properties and hints at novel ways to robustify and defend against adversarial attacks. We provide extensive empirical evidence on state-of-the-art network architectures to support our theoretical results.

[A Biologically Plausible Neural Network for Slow Feature Analysis](#)

- David Lipshutz, Charles Windolf, Siavash Golkar, Dmitri Chklovskii
- abstract@[open-review](#): Learning latent features from time series data is an important problem in both machine learning and brain function. One approach, called Slow Feature Analysis (SFA), leverages the slowness of many salient features relative to the rapidly varying input signals. Furthermore, when trained on naturalistic stimuli, SFA reproduces interesting properties of cells in the primary visual cortex and hippocampus, suggesting that the brain uses temporal slowness as a computational principle for learning latent features. However, despite the potential relevance of SFA for modeling brain function, there is currently no SFA algorithm with a biologically plausible neural network implementation, by which we mean an algorithm operates in the online setting and can be mapped onto a neural network with local synaptic updates. In this work, starting from an SFA objective, we derive an SFA algorithm, called Bio-SFA, with a biologically plausible neural network implementation. We validate Bio-SFA on naturalistic stimuli.

[Learning Feature Sparse Principal Subspace](#)

- Lai Tian, Feiping Nie, Rong Wang, Xuelong Li
- abstract@[open-review](#): This paper presents new algorithms to solve the feature-sparsity constrained PCA problem (FSPCA), which performs feature selection and PCA simultaneously. Existing optimization methods for FSPCA require data distribution assumptions and lack of global convergence guarantee. Though the general FSPCA problem is NP-hard, we show that, for a low-rank covariance, FSPCA can be solved globally (Algorithm 1). Then, we propose another strategy (Algorithm 2) to solve FSPCA for the general covariance by iteratively building a carefully designed proxy. We prove (data-dependent) approximation bound and convergence guarantees for the new algorithms. For the spectrum of covariance with exponential/Zipf's distribution, we provide exponential/posynomial approximation bound. Experimental results show the promising performance and efficiency of the new algorithms compared with the state-of-the-arts on both synthetic and real-world datasets.

[Online Adaptation for Consistent Mesh Reconstruction in the Wild](#)

- Xuetong Li, Sifei Liu, Shalini De Mello, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, Jan Kautz
- abstract@[open-review](#): This paper presents an algorithm to reconstruct temporally consistent 3D meshes of deformable object instances from videos in the wild. Without requiring annotations of 3D mesh, 2D keypoints, or camera pose for each video frame, we pose video-based reconstruction as a self-supervised online adaptation problem applied to any incoming test video. We first learn a category-specific 3D reconstruction model from a collection of single-view images of the same category that jointly predicts the shape, texture, and camera pose of an image. Then, at inference time, we adapt the model to a test video over time using self-supervised regularization terms that exploit temporal consistency of an object instance to enforce that all reconstructed meshes share a common texture map, a base shape, as well as parts. We demonstrate that our algorithm recovers temporally consistent and reliable 3D structures from videos of non-rigid objects including those of animals captured in the wild -- an extremely challenging task rarely addressed before.

[Online learning with dynamics: A minimax perspective](#)

- Kush Bhatia, Karthik Sridharan
- abstract@[open-review](#): We consider the problem of online learning with dynamics, where a learner interacts with a stateful environment over multiple rounds. In each round of the interaction, the learner selects a policy to deploy and incurs a cost that depends on both the chosen policy and current state of the world. The state-evolution dynamics and the costs are allowed to be time-varying, in a possibly adversarial way. In this setting, we study the problem of minimizing policy regret and provide non-constructive upper bounds on the minimax rate for the problem. Our main results provide sufficient conditions for online learnability for this setup with corresponding rates. The rates are characterized by: 1) a complexity term capturing the expressiveness of the underlying policy class under the dynamics of state change, and 2) a dynamics stability term measuring the deviation of the instantaneous loss from a certain counterfactual loss. Further, we provide matching lower bounds which show that both the complexity terms are indeed necessary. Our approach provides a unifying analysis that recovers regret bounds for several well studied problems including online learning with memory, online control of linear quadratic regulators, online Markov decision processes, and tracking adversarial targets. In addition, we show how our tools help obtain tight regret bounds for a new problems (with non-linear dynamics and non-convex losses) for which such bounds were not known prior to our work.

[Learning to Select Best Forecast Tasks for Clinical Outcome Prediction](#)

- Yuan Xue, Nan Du, Anne Mottram, Martin Seneviratne, Andrew M. Dai
- abstract@[open-review](#): The paradigm of pretraining' from a set of relevant auxiliary tasks and thenfinetuning' on a target task has been successfully applied in many different domains. However, when the auxiliary tasks are abundant, with complex relationships to the target task, using domain knowledge or searching over all possible pretraining setups are inefficient strategies. To address this challenge, we propose a method to automatically select from a large set of auxiliary tasks which yield a representation most useful to the target task. In particular, we develop an efficient algorithm that uses automatic auxiliary task selection within a nested-loop meta-learning process. We have applied this algorithm to the task of clinical outcome predictions in electronic medical records, learning from a large number of self-supervised tasks related to forecasting patient trajectories. Experiments on a real clinical dataset demonstrate the superior predictive performance of our method compared to direct supervised learning, naive pretraining and multitask learning, in particular in low-data scenarios when the primary task has very few examples. With detailed ablation analysis, we further show that the selection rules are interpretable and able to generalize to unseen target tasks with new data.

[Stochastic Optimization with Heavy-Tailed Noise via Accelerated Gradient Clipping](#)

- Eduard Gorbunov, Marina Danilova, Alexander Gasnikov
- abstract@[open-review](#): In this paper, we propose a new accelerated stochastic first-order method called clipped-SSTM for smooth convex stochastic optimization with heavy-tailed distributed noise in stochastic gradients and derive the first high-probability complexity bounds for this method closing the gap in the theory of stochastic optimization with heavy-tailed noise. Our method is based on a special variant of accelerated Stochastic Gradient Descent (SGD) and clipping of stochastic gradients. We extend our method to the strongly convex case and prove new complexity bounds that outperform state-of-the-art results in this case. Finally, we extend our proof technique and derive the first non-trivial high-probability complexity bounds for SGD with clipping without light-tails assumption on the noise.

[Adaptive Experimental Design with Temporal Interference: A Maximum Likelihood Approach](#)

- Peter W. Glynn, Ramesh Johari, Mohammad Rasouli
- abstract@[open-review](#): Suppose an online platform wants to compare a treatment and control policy (e.g., two different matching algorithms in a ridesharing system, or two different inventory management algorithms in an online retail site). Standard experimental approaches to this problem are biased (due to temporal interference between the policies), and not sample efficient. We study optimal experimental design for this setting. We view testing the two policies as the problem of estimating the steady state difference in reward between two unknown Markov chains (i.e., policies). We assume estimation of the steady state reward for each chain proceeds via nonparametric maximum likelihood, and search for consistent (i.e., asymptotically unbiased) experimental designs that are efficient (i.e., asymptotically minimum variance). Characterizing such designs is equivalent to a Markov decision problem with a minimum variance objective; such problems generally do not admit tractable solutions. Remarkably, in our setting, using a novel application of classical martingale analysis of Markov chains via Poisson's equation, we characterize efficient designs via a succinct convex optimization problem. We use this characterization to propose a consistent, efficient online experimental design that adaptively samples the two Markov chains.

[From Trees to Continuous Embeddings and Back: Hyperbolic Hierarchical Clustering](#)

- Ines Chami, Albert Gu, Vaggos Chatziafratis, Christopher RÃ©
- abstract@[open-review](#): Similarity-based Hierarchical Clustering (HC) is a classical unsupervised machine learning algorithm that has traditionally been solved with heuristic algorithms like Average-Linkage. Recently, Dasgupta reframed HC as a discrete optimization problem by introducing a global cost function measuring the quality of a given tree. In this work, we provide the first continuous relaxation of Dasgupta's discrete optimization problem with provable quality guarantees. The key idea of our method, HypHC, is showing a direct correspondence from discrete trees to continuous representations (via the hyperbolic embeddings of their leaf nodes) and back (via a decoding algorithm that maps leaf embeddings to a dendrogram), allowing us to search the space of discrete binary trees with continuous optimization. Building on analogies between trees and hyperbolic space, we derive a continuous analogue for the notion of lowest common ancestor, which leads to a continuous relaxation of Dasgupta's discrete objective. We can show that after decoding, the global minimizer of our continuous relaxation yields a discrete tree with a $(1+\epsilon)$ -factor approximation for Dasgupta's optimal tree, where ϵ can be made arbitrarily small and controls optimization challenges. We experimentally evaluate HypHC on a variety of HC benchmarks and find that even approximate solutions found with gradient descent have superior clustering quality than agglomerative heuristics or other gradient based algorithms. Finally, we highlight the flexibility of HypHC using end-to-end training in a downstream classification task.

[The Autoencoding Variational Autoencoder](#)

- Taylan Cemgil, Sumedh Ghaisas, Krishnamurthy Dvijotham, Sven Gowal, Pushmeet Kohli
- abstract@[open-review](#): Does a Variational AutoEncoder (VAE) consistently encode typical samples generated from its decoder? This paper shows that the perhaps surprising answer to this question is 'No'; a (nominally trained) VAE does not necessarily amortize inference for typical samples that it is capable of generating. We study the implications of this behaviour on the learned representations and also the consequences of fixing it by introducing a notion of self consistency. Our approach hinges on an alternative construction of the variational approximation distribution to the true posterior of an extended VAE model with a Markov chain alternating between the encoder and the decoder. The method can be used to train a VAE model from scratch or given an already trained VAE, it can be run as a post processing step in an entirely self supervised way without access to the original training data. Our experimental analysis reveals that encoders trained with our self-consistency approach lead to representations that are robust (insensitive) to perturbations in the input introduced by adversarial attacks. We provide experimental results on the ColorMnist and CelebA benchmark datasets that quantify the properties of the learned representations and compare the approach with a baseline that is specifically trained for the desired property.

[A Fair Classifier Using Kernel Density Estimation](#)

- Jaewoong Cho, Gyeongjo Hwang, Changho Suh

- abstract@[open-review](#): As machine learning becomes prevalent in a widening array of sensitive applications such as job hiring and criminal justice, one critical aspect that machine learning classifiers should respect is to ensure fairness: guaranteeing the irrelevancy of a prediction output to sensitive attributes such as gender and race. In this work, we develop a kernel density estimation trick to quantify fairness measures that capture the degree of the irrelevancy. A key feature of our approach is that quantified fairness measures can be expressed as differentiable functions w.r.t. classifier model parameters. This then allows us to enjoy prominent gradient descent to readily solve an interested optimization problem that fully respects fairness constraints. We focus on a binary classification setting and two well-known definitions of group fairness: Demographic Parity (DP) and Equalized Odds (EO). Our experiments both on synthetic and benchmark real datasets demonstrate that our algorithm outperforms prior fair classifiers in accuracy-fairness tradeoff performance both w.r.t. DP and EO.

[A Randomized Algorithm to Reduce the Support of Discrete Measures](#)

- Francesco Cosentino, Harald Oberhauser, Alessandro Abate
- abstract@[open-review](#): Given a discrete probability measure supported on $\$N\$$ atoms and a set of $\$n\$$ real-valued functions, there exists a probability measure that is supported on a subset of $\$n+1\$$ of the original $\$N\$$ atoms and has the same mean when integrated against each of the $\$n\$$ functions. If $\$ N \gg n \$$ this results in a huge reduction of complexity. We give a simple geometric characterization of barycenters via negative cones and derive a randomized algorithm that computes this new measure by ``greedy geometric sampling''. We then study its properties, and benchmark it on synthetic and real-world data to show that it can be very beneficial in the $\$N \gg n \$$ regime. A Python implementation is available at [url{https://github.com/FraCose/Recombination_Random_Algos}](https://github.com/FraCose/Recombination_Random_Algos).

[Distributionally Robust Federated Averaging](#)

- Yuyang Deng, Mohammad Mahdi Kamani, Mehrdad Mahdavi
- abstract@[open-review](#): In this paper, we study communication efficient distributed algorithms for distributionally robust federated learning via periodic averaging with adaptive sampling. In contrast to standard empirical risk minimization, due to the minimax structure of the underlying optimization problem, a key difficulty arises from the fact that the global parameter that controls the mixture of local losses can only be updated infrequently on the global stage. To compensate for this, we propose a Distributionally Robust Federated Averaging (DRFA) algorithm that employs a novel snapshotting scheme to approximate the accumulation of history gradients of the mixing parameter. We analyze the convergence rate of DRFA in both convex-linear and nonconvex-linear settings. We also generalize the proposed idea to objectives with regularization on the mixture parameter and propose a proximal variant, dubbed as DRFA-Prox, with provable convergence rates. We also analyze an alternative optimization method for regularized case in strongly-convex-strongly-concave and non-convex (under PL condition)-strongly-concave settings. To the best of our knowledge, this paper is the first to solve distributionally robust federated learning with reduced communication, and to analyze the efficiency of local descent methods on distributed minimax problems. We give corroborating experimental evidence for our theoretical results in federated learning settings.

[Sharp uniform convergence bounds through empirical centralization](#)

- Cyrus Cousins, Matteo Riondato
- abstract@[open-review](#): We introduce the use of empirical centralization to derive novel practical, probabilistic, sample-dependent bounds to the Supremum Deviation (SD) of empirical means of functions in a family from their expectations. Our bounds have optimal dependence on the maximum (i.e., wimpy) variance and the function ranges, and the same dependence on the number of samples as existing SD bounds. To compute the SD bounds in practice, we develop tightly-concentrated Monte Carlo estimators of the empirical Rademacher average of the empirically-centralized family, and we show novel concentration results for the empirical wimpy variance. Our experimental evaluation shows that our bounds greatly outperform non-centralized bounds and are extremely practical even at small sample sizes.

[COBE: Contextualized Object Embeddings from Narrated Instructional Video](#)

- Gedas Bertasius, Lorenzo Torresani
- abstract@[open-review](#): Many objects in the real world undergo dramatic variations in visual appearance. For example, a tomato may be red or green, sliced or chopped, fresh or fried, liquid or solid. Training a single detector to accurately recognize tomatoes in all these different states is challenging. On the other hand, contextual cues (e.g., the presence of a knife, a cutting board, a strainer or a pan) are often strongly indicative of how the object appears in the scene. Recognizing such contextual cues is useful not only to improve the accuracy of object detection or to determine the state of the object, but also to understand its functional properties and to infer ongoing or upcoming human-object interactions. A fully-supervised approach to recognizing object states and their contexts in the real-world is unfortunately marred by the long-tailed, open-ended distribution of the data, which would effectively require massive amounts of annotations to capture the appearance of objects in all their different forms. Instead of relying on manually-labeled data for this task, we propose a new framework for learning Contextualized OObject Embeddings (COBE) from automatically-transcribed narrations of instructional videos. We leverage the semantic and compositional structure of language by training a visual detector to predict a contextualized word embedding of the object and its associated narration. This enables the learning of an object representation where concepts relate according to a semantic language metric. Our experiments show that our detector learns to predict a rich variety of contextual object information, and that it is highly effective in the settings of few-shot and zero-shot learning.

[Knowledge Transfer in Multi-Task Deep Reinforcement Learning for Continuous Control](#)

- Zhiyuan Xu, Kun Wu, Zhengping Che, Jian Tang, Jieping Ye
- abstract@[open-review](#): While Deep Reinforcement Learning (DRL) has emerged as a promising approach to many complex tasks, it remains challenging to train a single DRL agent that is capable of undertaking multiple different continuous control tasks. In this paper, we present a Knowledge Transfer based Multi-task Deep Reinforcement Learning framework (KTM-DRL) for continuous control, which enables a single DRL agent to achieve expert-level performance in multiple different tasks by learning from task-specific teachers. In KTM-DRL, the multi-task agent first leverages an offline knowledge transfer algorithm designed particularly for the actor-critic architecture to quickly learn a control policy from the experience of task-specific teachers, and then it employs an online learning algorithm to further improve itself by learning from new online transition samples under the guidance of those teachers. We perform a comprehensive empirical study with two commonly-used benchmarks in the MuJoCo continuous control task suite. The experimental results well justify the effectiveness of KTM-DRL and its knowledge transfer and online learning algorithms, as well as its superiority over the state-of-the-art by a large margin.

[Finite Versus Infinite Neural Networks: an Empirical Study](#)

- Jaehoon Lee, Samuel Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, Jascha Sohl-Dickstein
- abstract@[open-review](#): We perform a careful, thorough, and large scale empirical study of the correspondence between wide neural networks and kernel methods. By doing so, we resolve a variety of open questions related to the study of infinitely wide neural networks. Our experimental results include: kernel methods outperform fully-connected finite-width networks, but underperform convolutional finite width networks; neural network Gaussian process (NNGP) kernels frequently outperform neural tangent (NT) kernels; centered and ensembled finite networks have reduced posterior variance and behave more similarly to infinite networks; weight decay and the use of a large learning rate break the correspondence between finite and infinite networks; the NTK parameterization outperforms the standard parameterization for finite width networks; diagonal regularization of kernels acts similarly to early stopping; floating point precision limits kernel performance beyond a critical dataset size; regularized ZCA whitening improves accuracy; finite network

performance depends non-monotonically on width in ways not captured by double descent phenomena; equivariance of CNNs is only beneficial for narrow networks far from the kernel regime. Our experiments additionally motivate an improved layer-wise scaling for weight decay which improves generalization in finite-width networks. Finally, we develop improved best practices for using NNGP and NT kernels for prediction, including a novel ensembling technique. Using these best practices we achieve state-of-the-art results on CIFAR-10 classification for kernels corresponding to each architecture class we consider.

[Supermasks in Superposition](#)

- Mitchell Wortsman, Vivek Ramanujan, Rosanne Liu, Aniruddha Kembhavi, Mohammad Rastegari, Jason Yosinski, Ali Farhadi
- abstract@[open-review](#): We present the Supermasks in Superposition (SupSup) model, capable of sequentially learning thousands of tasks without catastrophic forgetting. Our approach uses a randomly initialized, fixed base network and for each task finds a subnetwork (supermask) that achieves good performance. If task identity is given at test time, the correct subnetwork can be retrieved with minimal memory usage. If not provided, SupSup can infer the task using gradient-based optimization to find a linear superposition of learned supermasks which minimizes the output entropy. In practice we find that a single gradient step is often sufficient to identify the correct mask, even among 2500 tasks. We also showcase two promising extensions. First, SupSup models can be trained entirely without task identity information, as they may detect when they are uncertain about new data and allocate an additional supermask for the new training distribution. Finally the entire, growing set of supermasks can be stored in a constant-sized reservoir by implicitly storing them as attractors in a fixed-sized Hopfield network.

[Nonasymptotic Guarantees for Spiked Matrix Recovery with Generative Priors](#)

- Jorio Cocola, Paul Hand, Vlad Voroninski
- abstract@[open-review](#): Many problems in statistics and machine learning require the reconstruction of a rank-one signal matrix from noisy data. Enforcing additional prior information on the rank-one component is often key to guaranteeing good recovery performance. One such prior on the low-rank component is sparsity, giving rise to the sparse principal component analysis problem. Unfortunately, there is strong evidence that this problem suffers from a computational-to-statistical gap, which may be fundamental. In this work, we study an alternative prior where the low-rank component is in the range of a trained generative network. We provide a non-asymptotic analysis with optimal sample complexity, up to logarithmic factors, for rank-one matrix recovery under an expansive-Gaussian network prior. Specifically, we establish a favorable global optimization landscape for a nonlinear least squares objective, provided the number of samples is on the order of the dimensionality of the input to the generative model. This result suggests that generative priors have no computational-to-statistical gap for structured rank-one matrix recovery in the finite data, nonasymptotic regime. We present this analysis in the case of both the Wishart and Wigner spiked matrix models.

[Almost Optimal Model-Free Reinforcement Learning via Reference-Advantage Decomposition](#)

- Zihan Zhang, Yuan Zhou, Xiangyang Ji
- abstract@[open-review](#): We study the reinforcement learning problem in the setting of finite-horizon episodic Markov Decision Processes (MDPs) with S states, A actions, and episode length H . We propose a model-free algorithm UCB-ADVANTAGE and prove that it achieves $\tilde{O}(\sqrt{H^2 SAT})$ regret where $T=KH$ and K is the number of episodes to play. Our regret bound improves upon the results of [Jin et al., 2018] and matches the best known model-based algorithms as well as the information theoretic lower bound up to logarithmic factors. We also show that UCB-ADVANTAGE achieves low local switching cost and applies to concurrent reinforcement learning, improving upon the recent results of [Bai et al., 2019].

[Learning to Incentivize Other Learning Agents](#)

- Jiachen Yang, Ang Li, Mehrdad Farajtabar, Peter Sunehag, Edward Hughes, Hongyuan Zha
- abstract@[open-review](#): The challenge of developing powerful and general Reinforcement Learning (RL) agents has received increasing attention in recent years. Much of this effort has focused on the single-agent setting, in which an agent maximizes a predefined extrinsic reward function. However, a long-term question inevitably arises: how will such independent agents cooperate when they are continually learning and acting in a shared multi-agent environment? Observing that humans often provide incentives to influence others' behavior, we propose to equip each RL agent in a multi-agent environment with the ability to give rewards directly to other agents, using a learned incentive function. Each agent learns its own incentive function by explicitly accounting for its impact on the learning of recipients and, through them, the impact on its own extrinsic objective. We demonstrate in experiments that such agents significantly outperform standard RL and opponent-shaping agents in challenging general-sum Markov games, often by finding a near-optimal division of labor. Our work points toward more opportunities and challenges along the path to ensure the common good in a multi-agent future.

[Displacement-Invariant Matching Cost Learning for Accurate Optical Flow Estimation](#)

- Jianyuan Wang, Yiran Zhong, Yuchao Dai, Kaihao Zhang, Pan Ji, Hongdong Li
- abstract@[open-review](#): Learning matching costs has been shown to be critical to the success of the state-of-the-art deep stereo matching methods, in which 3D convolutions are applied on a 4D feature volume to learn a 3D cost volume. However, this mechanism has never been employed for the optical flow task. This is mainly due to the significantly increased search dimension in the case of optical flow computation, i.e., a straightforward extension would require dense 4D convolutions in order to process a 5D feature volume, which is computationally prohibitive. This paper proposes a novel solution that is able to bypass the requirement of building a 5D feature volume while still allowing the network to learn suitable matching costs from data. Our key innovation is to decouple the connection between 2D displacements and learn the matching costs at each 2D displacement hypothesis independently, i.e., displacement-invariant cost learning. Specifically, we apply the same 2D convolution-based matching net independently on each 2D displacement hypothesis to learn a 4D cost volume. Moreover, we propose a displacement-aware projection layer to scale the learned cost volume, which reconsiders the correlation between different displacement candidates and mitigates the multi-modal problem in the learned cost volume. The cost volume is then projected to optical flow estimation through a 2D soft-argmin layer. Extensive experiments show that our approach achieves state-of-the-art accuracy on various datasets, and outperforms all published optical flow methods on the Sintel benchmark. The code is available at <https://github.com/jytime/DICL-Flow>.

[Distributionally Robust Local Non-parametric Conditional Estimation](#)

- Viet Anh Nguyen, Fan Zhang, Jose Blanchet, Erick Delage, Yinyu Ye
- abstract@[open-review](#): Conditional estimation given specific covariate values (i.e., local conditional estimation or functional estimation) is ubiquitously useful with applications in engineering, social and natural sciences. Existing data-driven non-parametric estimators mostly focus on structured homogeneous data (e.g., weakly independently and stationary data), thus they are sensitive to adversarial noise and may perform poorly under a low sample size. To alleviate these issues, we propose a new distributionally robust estimator that generates non-parametric local estimates by minimizing the worst-case conditional expected loss over all adversarial distributions in a Wasserstein ambiguity set. We show that despite being generally intractable, the local estimator can be efficiently found via convex optimization under broadly applicable settings, and it is robust to the corruption and heterogeneity of the data. Various experiments show the competitive performance of this new class of estimator.

[Robust Multi-Object Matching via Iterative Reweighting of the Graph Connection Laplacian](#)

- Yunpeng Shi, Shaohan Li, Gilad Lerman
- abstract@[open-review](#): We propose an efficient and robust iterative solution to the multi-object matching problem. We first clarify serious limitations of current methods as well as the inappropriateness of the standard iteratively reweighted least squares procedure. In view of these limitations, we suggest a novel and more reliable iterative reweighting strategy that incorporates information from higher-order neighborhoods by exploiting the graph connection Laplacian. We demonstrate the superior performance of our procedure over state-of-the-art methods using both synthetic and real datasets.

[Meta-Gradient Reinforcement Learning with an Objective Discovered Online](#)

- Zhongwen Xu, Hado P. van Hasselt, Matteo Hessel, Junhyuk Oh, Satinder Singh, David Silver
- abstract@[open-review](#): Deep reinforcement learning includes a broad family of algorithms that parameterise an internal representation, such as a value function or policy, by a deep neural network. Each algorithm optimises its parameters with respect to an objective, such as Q-learning or policy gradient, that defines its semantics. In this work, we propose an algorithm based on meta-gradient descent that discovers its own objective, flexibly parameterised by a deep neural network, solely from interactive experience with its environment. Over time, this allows the agent to learn how to learn increasingly effectively. Furthermore, because the objective is discovered online, it can adapt to changes over time. We demonstrate that the algorithm discovers how to address several important issues in RL, such as bootstrapping, non-stationarity, and off-policy learning. On the Atari Learning Environment, the meta-gradient algorithm adapts over time to learn with greater efficiency, eventually outperforming the median score of a strong actor-critic baseline.

[Learning Strategy-Aware Linear Classifiers](#)

- Yiling Chen, Yang Liu, Chara Podimata
- abstract@[open-review](#): We address the question of repeatedly learning linear classifiers against agents who are \emph{strategically} trying to \emph{game} the deployed classifiers, and we use the \emph{Stackelberg regret} to measure the performance of our algorithms. First, we show that Stackelberg and external regret for the problem of strategic classification are \emph{strongly incompatible}: i.e., there exist worst-case scenarios, where \emph{any} sequence of actions providing \emph{sublinear} external regret might result in \emph{linear} Stackelberg regret and vice versa. Second, we present a strategy-aware algorithm for minimizing the Stackelberg regret for which we prove nearly matching upper and lower regret bounds. Finally, we provide simulations to complement our theoretical analysis. Our results advance the growing literature of learning from revealed preferences, which has so far focused on ``smoother'' assumptions from the perspective of the learner and the agents respectively.

[Upper Confidence Primal-Dual Reinforcement Learning for CMDP with Adversarial Loss](#)

- Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, Zhaoran Wang
- abstract@[open-review](#): We consider online learning for episodic stochastically constrained Markov decision processes (CMDP), which plays a central role in ensuring the safety of reinforcement learning. Here the loss function can vary arbitrarily across the episodes, whereas both the loss received and the budget consumption are revealed at the end of each episode. Previous works solve this problem under the restrictive assumption that the transition model of the MDP is known a priori and establish regret bounds that depend polynomially on the cardinalities of the state space \mathcal{S} and the action space \mathcal{A} . In this work, we propose a new \emph{upper confidence primal-dual} algorithm, which only requires the trajectories sampled from the transition model. In particular, we prove that the proposed algorithm achieves $\widetilde{O}(L\sqrt{\mathcal{S}\mathcal{A}T})$ upper bounds of both the regret and the constraint violation, where L is the length of each episode. Our analysis incorporates a new high-probability drift analysis of Lagrange multiplier processes into the celebrated regret analysis of upper confidence reinforcement learning, which demonstrates the power of ``optimism in the face of uncertainty'' in constrained online learning.

[Calibrating Deep Neural Networks using Focal Loss](#)

- Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, Puneet Dokania
- abstract@[open-review](#): Miscalibration -- a mismatch between a model's confidence and its correctness -- of Deep Neural Networks (DNNs) makes their predictions hard to rely on. Ideally, we want networks to be accurate, calibrated and confident. We show that, as opposed to the standard cross-entropy loss, focal loss (Lin et al., 2017) allows us to learn models that are already very well calibrated. When combined with temperature scaling, whilst preserving accuracy, it yields state-of-the-art calibrated models. We provide a thorough analysis of the factors causing miscalibration, and use the insights we glean from this to justify the empirically excellent performance of focal loss. To facilitate the use of focal loss in practice, we also provide a principled approach to automatically select the hyperparameter involved in the loss function. We perform extensive experiments on a variety of computer vision and NLP datasets, and with a wide variety of network architectures, and show that our approach achieves state-of-the-art calibration without compromising on accuracy in almost all cases. Code is available at https://github.com/torrvision/focal_calibration.

[Optimizing Mode Connectivity via Neuron Alignment](#)

- Norman Tatro, Pin-Yu Chen, Payel Das, Igor Melnyk, Prasanna Sattigeri, Rongjie Lai
- abstract@[open-review](#): The loss landscapes of deep neural networks are not well understood due to their high nonconvexity. Empirically, the local minima of these loss functions can be connected by a learned curve in model space, along which the loss remains nearly constant; a feature known as mode connectivity. Yet, current curve finding algorithms do not consider the influence of symmetry in the loss surface created by model weight permutations. We propose a more general framework to investigate the effect of symmetry on landscape connectivity by accounting for the weight permutations of the networks being connected. To approximate the optimal permutation, we introduce an inexpensive heuristic referred to as neuron alignment. Neuron alignment promotes similarity between the distribution of intermediate activations of a model along the curve with that of the endpoint models. We provide theoretical analysis establishing the benefit of alignment to mode connectivity based on this simple heuristic. We empirically verify that the permutation given by alignment is locally optimal via a proximal alternating minimization scheme. Empirically, optimizing the weight permutation is critical for efficiently learning a simple, planar, low-loss curve between networks that successfully generalizes. Our alignment method can significantly alleviate the recently identified robust loss barrier on the path connecting two adversarial robust models and find more robust and accurate models on the path.

[Information Theoretic Regret Bounds for Online Nonlinear Control](#)

- Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, Wen Sun
- abstract@[open-review](#): This work studies the problem of sequential control in an unknown, nonlinear dynamical system, where we model the underlying system dynamics as an unknown function in a known Reproducing Kernel Hilbert Space. This framework yields a general setting that permits discrete and continuous control inputs as well as non-smooth, non-differentiable dynamics. Our main result, the Lower Confidence-based Continuous Control (LC3) algorithm, enjoys a near-optimal $O(\sqrt{T})$ regret bound against the optimal controller in episodic settings, where T is the number of episodes. The bound has no explicit dependence on dimension of the system dynamics, which could be infinite, but instead only depends on information theoretic quantities. We empirically show its application to a number of nonlinear control tasks and demonstrate the benefit of exploration for learning model dynamics.

[A kernel test for quasi-independence](#)

- Tamara Fernandez, Wenkai Xu, Marc Ditzhaus, Arthur Gretton
- abstract@[open-review](#): We consider settings in which the data of interest correspond to pairs of ordered times, e.g., the birth times of the first and second child, the times at which a new user creates an account and makes the first purchase on a website, and the entry and survival times of patients in a clinical trial. In these settings, the two times are not independent (the second occurs after the first), yet it is still of interest to determine whether there exists significant dependence "beyond" their ordering in time. We refer to this notion as "quasi-(in)dependence." For instance, in a clinical trial, to avoid biased selection, we might wish to verify that recruitment times are quasi-independent of survival times, where dependencies might arise due to seasonal effects. In this paper, we propose a nonparametric statistical test of quasi-independence. Our test considers a potentially infinite space of alternatives, making it suitable for complex data where the nature of the possible quasi-dependence is not known in advance. Standard parametric approaches are recovered as special cases, such as the classical conditional Kendall's tau, and log-rank tests. The tests apply in the right-censored setting: an essential feature in clinical trials, where patients can withdraw from the study. We provide an asymptotic analysis of our test-statistic, and demonstrate in experiments that our test obtains better power than existing approaches, while being more computationally efficient.

[First Order Constrained Optimization in Policy Space](#)

- Yiming Zhang, Quan Vuong, Keith Ross
- abstract@[open-review](#): In reinforcement learning, an agent attempts to learn high-performing behaviors through interacting with the environment, such behaviors are often quantified in the form of a reward function. However some aspects of behavior—such as ones which are deemed unsafe and to be avoided—are best captured through constraints. We propose a novel approach called First Order Constrained Optimization in Policy Space (FOCOPS) which maximizes an agent's overall reward while ensuring the agent satisfies a set of cost constraints. Using data generated from the current policy, FOCOPS first finds the optimal update policy by solving a constrained optimization problem in the nonparameterized policy space. FOCOPS then projects the update policy back into the parametric policy space. Our approach has an approximate upper bound for worst-case constraint violation throughout training and is first-order in nature therefore simple to implement. We provide empirical evidence that our simple approach achieves better performance on a set of constrained robotics locomotive tasks.

[Learning Augmented Energy Minimization via Speed Scaling](#)

- Etienne Bamas, Andreas Maggiori, Lars Rohwedder, Ola Svensson
- abstract@[open-review](#): As power management has become a primary concern in modern data centers, computing resources are being scaled dynamically to minimize energy consumption. We initiate the study of a variant of the classic online speed scaling problem, in which machine learning predictions about the future can be integrated naturally. Inspired by recent work on learning-augmented online algorithms, we propose an algorithm which incorporates predictions in a black-box manner and outperforms any online algorithm if the accuracy is high, yet maintains provable guarantees if the prediction is very inaccurate. We provide both theoretical and experimental evidence to support our claims.

[Exploiting MMD and Sinkhorn Divergences for Fair and Transferable Representation Learning](#)

- Luca Oneto, Michele Donini, Giulia Luise, Carlo Ciliberto, Andreas Maurer, Massimiliano Pontil
- abstract@[open-review](#): Developing learning methods which do not discriminate subgroups in the population is a central goal of algorithmic fairness. One way to reach this goal is by modifying the data representation in order to meet certain fairness constraints. In this work we measure fairness according to demographic parity. This requires the probability of the possible model decisions to be independent of the sensitive information. We argue that the goal of imposing demographic parity can be substantially facilitated within a multitask learning setting. We present a method for learning a shared fair representation across multiple tasks, by means of different new constraints based on MMD and Sinkhorn Divergences. We derive learning bounds establishing that the learned representation transfers well to novel tasks. We present experiments on three real world datasets, showing that the proposed method outperforms state-of-the-art approaches by a significant margin.

[Deep Rao-Blackwellised Particle Filters for Time Series Forecasting](#)

- Richard Kurle, Syama Sundar Rangapuram, Emmanuel de Bézenac, Stephan Günnemann, Jan Gasthaus
- abstract@[open-review](#): This work addresses efficient inference and learning in switching Gaussian linear dynamical systems using a Rao-Blackwellised particle filter and a corresponding Monte Carlo objective. To improve the forecasting capabilities, we extend this classical model by conditionally linear state-to-switch dynamics, while leaving the partial tractability of the conditional Gaussian linear part intact. Furthermore, we use an auxiliary variable approach with a decoder-type neural network that allows for more complex non-linear emission models and multivariate observations. We propose a Monte Carlo objective that leverages the conditional linearity by computing the corresponding conditional expectations in closed-form and a suitable proposal distribution that is factorised similarly to the optimal proposal distribution. We evaluate our approach on several popular time series forecasting datasets as well as image streams of simulated physical systems. Our results show improved forecasting performance compared to other deep state-space model approaches.

[Why are Adaptive Methods Good for Attention Models?](#)

- Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, Suvrit Sra
- abstract@[open-review](#): While stochastic gradient descent (SGD) is still the de facto algorithm in deep learning, adaptive methods like Clipped SGD/Adam have been observed to outperform SGD across important tasks, such as attention models. The settings under which SGD performs poorly in comparison to adaptive methods are not well understood yet. In this paper, we provide empirical and theoretical evidence that a heavy-tailed distribution of the noise in stochastic gradients is one cause of SGD's poor performance. We provide the first tight upper and lower convergence bounds for adaptive gradient methods under heavy-tailed noise. Further, we demonstrate how gradient clipping plays a key role in addressing heavy-tailed gradient noise. Subsequently, we show how clipping can be applied in practice by developing an adaptive coordinate-wise clipping algorithm (ACClip) and demonstrate its superior performance on BERT pretraining and finetuning tasks.

[Neural Sparse Representation for Image Restoration](#)

- Yuchen Fan, Jiahui Yu, Yiqun Mei, Yulun Zhang, Yun Fu, Ding Liu, Thomas S. Huang
- abstract@[open-review](#): Inspired by the robustness and efficiency of sparse representation in sparse coding based image restoration models, we investigate the sparsity of neurons in deep networks. Our method structurally enforces sparsity constraints upon hidden neurons. The sparsity constraints are favorable for gradient-based learning algorithms and attachable to convolution layers in various networks. Sparsity in neurons enables computation saving by only operating on non-zero components without hurting accuracy. Meanwhile, our method can magnify representation dimensionality and model capacity with negligible additional computation cost. Experiments show that sparse representation is crucial in deep neural networks for multiple image restoration tasks, including image super-resolution, image denoising, and image compression artifacts removal.

[Boosting First-Order Methods by Shifting Objective: New Schemes with Faster Worst-Case Rates](#)

- Kaiwen Zhou, Anthony Man-Cho So, James Cheng

- abstract@[open-review](#): We propose a new methodology to design first-order methods for unconstrained strongly convex problems. Specifically, instead of tackling the original objective directly, we construct a shifted objective function that has the same minimizer as the original objective and encodes both the smoothness and strong convexity of the original objective in an interpolation condition. We then propose an algorithmic template for tackling the shifted objective, which can exploit such a condition. Following this template, we derive several new accelerated schemes for problems that are equipped with various first-order oracles and show that the interpolation condition allows us to vastly simplify and tighten the analysis of the derived methods. In particular, all the derived methods have faster worst-case convergence rates than their existing counterparts. Experiments on machine learning tasks are conducted to evaluate the new methods.

[Robust Sequence Submodular Maximization](#)

- Gamal Sallam, Zizhan Zheng, Jie Wu, Bo Ji
- abstract@[open-review](#): Submodularity is an important property of set functions and has been extensively studied in the literature. It models set functions that exhibit a diminishing returns property, where the marginal value of adding an element to a set decreases as the set expands. This notion has been generalized to considering sequence functions, where the order of adding elements plays a crucial role and determines the function value; the generalized notion is called sequence (or string) submodularity. In this paper, we study a new problem of robust sequence submodular maximization with cardinality constraints. The robustness is against the removal of a subset of elements in the selected sequence (e.g., due to malfunctions or adversarial attacks). Compared to robust submodular maximization for set function, new challenges arise when sequence functions are concerned. Specifically, there are multiple definitions of submodularity for sequence functions, which exhibit subtle yet critical differences. Another challenge comes from two directions of monotonicity: forward monotonicity and backward monotonicity, both of which are important to proving performance guarantees. To address these unique challenges, we design two robust greedy algorithms: while one algorithm achieves a constant approximation ratio but is robust only against the removal of a subset of contiguous elements, the other is robust against the removal of an arbitrary subset of the selected elements but requires a stronger assumption and achieves an approximation ratio that depends on the number of the removed elements. Finally, we generalize the analyses to considering sequence functions under weaker assumptions based on approximate versions of sequence submodularity and backward monotonicity.

[Certified Monotonic Neural Networks](#)

- Xingchao Liu, Xing Han, Na Zhang, Qiang Liu
- abstract@[open-review](#): Learning monotonic models with respect to a subset of the inputs is a desirable feature to effectively address the fairness, interpretability, and generalization issues in practice. Existing methods for learning monotonic neural networks either require specifically designed model structures to ensure monotonicity, which can be too restrictive/complicated, or enforce monotonicity by adjusting the learning process, which cannot provably guarantee the learned model is monotonic on selected features. In this work, we propose to certify the monotonicity of the general piece-wise linear neural networks by solving a mixed integer linear programming problem. This provides a new general approach for learning monotonic neural networks with arbitrary model structures. Our method allows us to train neural networks with heuristic monotonicity regularizations, and we can gradually increase the regularization magnitude until the learned network is certified monotonic. Compared to prior work, our method does not require human-designed constraints on the weight space and also yields more accurate approximation. Empirical studies on various datasets demonstrate the efficiency of our approach over the state-of-the-art methods, such as Deep Lattice Networks

[System Identification with Biophysical Constraints: A Circuit Model of the Inner Retina](#)

- Cornelius Schräder, David Klindt, Sarah Strauss, Katrin Franke, Matthias Bethge, Thomas Euler, Philipp Berens
- abstract@[open-review](#): Visual processing in the retina has been studied in great detail at all levels such that a comprehensive picture of the retina's cell types and the many neural circuits they form is emerging. However, the currently best performing models of retinal function are black-box CNN models which are agnostic to such biological knowledge. In particular, these models typically neglect the role of the many inhibitory circuits involving amacrine cells and the biophysical mechanisms underlying synaptic release. Here, we present a computational model of temporal processing in the inner retina, including inhibitory feedback circuits and realistic synaptic release mechanisms. Fit to the responses of bipolar cells, the model generalized well to new stimuli including natural movie sequences, performing on par with or better than a benchmark black-box model. In pharmacology experiments, the model replicated in silico the effect of blocking specific amacrine cell populations with high fidelity, indicating that it had learned key circuit functions. Also, more in depth comparisons showed that connectivity patterns learned by the model were well matched to connectivity patterns extracted from connectomics data. Thus, our model provides a biologically interpretable data-driven account of temporal processing in the inner retina, filling the gap between purely black-box and detailed biophysical modeling.

[Efficient Algorithms for Device Placement of DNN Graph Operators](#)

- Jakub M. Tarnawski, Amar Phanishayee, Nikhil Devanur, Divya Mahajan, Fanny Nina Paravecino
- abstract@[open-review](#): Modern machine learning workloads use large models, with complex structures, that are very expensive to execute. The devices that execute complex models are becoming increasingly heterogeneous as we see a flourishing of Domain Specific Architectures (DSAs) being offered as hardware accelerators in addition to CPUs. These trends necessitate distributing the workload across multiple devices. Recent work has shown that significant gains can be obtained with model parallelism, i.e., partitioning a neural network's computational graph onto multiple devices. In particular, this form of parallelism assumes a pipeline of devices, which is fed a stream of samples and yields high throughput for training and inference of DNNs. However, for such settings (large models and multiple heterogeneous devices), we require automated algorithms and toolchains that can partition the ML workload across devices. In this paper, we identify and isolate the structured optimization problem at the core of device placement of DNN operators, for both inference and training, especially in modern pipelined settings. We then provide algorithms that solve this problem to optimality. We demonstrate the applicability and efficiency of our approaches using several contemporary DNN computation graphs.

[Active Invariant Causal Prediction: Experiment Selection through Stability](#)

- Juan L. Gamella, Christina Heinze-Deml
- abstract@[open-review](#): A fundamental difficulty of causal learning is that causal models can generally not be fully identified based on observational data only. Interventional data, that is, data originating from different experimental environments, improves identifiability. However, the improvement depends critically on the target and nature of the interventions carried out in each experiment. Since in real applications experiments tend to be costly, there is a need to perform the right interventions such that as few as possible are required. In this work we propose a new active learning (i.e. experiment selection) framework (A-ICP) based on Invariant Causal Prediction (ICP) (Peters et al. 2016). For general structural causal models, we characterize the effect of interventions on so-called stable sets, a notion introduced by Pfister et al. 2019. We leverage these results to propose several intervention selection policies for A-ICP which quickly reveal the direct causes of a response variable in the causal graph while maintaining the error control inherent in ICP. Empirically, we analyze the performance of the proposed policies in both population and finite-regime experiments.

[BOSS: Bayesian Optimization over String Spaces](#)

- Henry Moss, David Leslie, Daniel Beck, Javier González, Paul Rayson
- abstract@[open-review](#): This article develops a Bayesian optimization (BO) method which acts directly over raw strings, proposing the first uses of string kernels and genetic algorithms within BO loops. Recent applications of BO over strings have been hindered by the need to map inputs into a smooth and unconstrained latent space. Learning this projection is computationally and data-intensive. Our approach instead builds a powerful Gaussian process

surrogate model based on string kernels, naturally supporting variable length inputs, and performs efficient acquisition function maximization for spaces with syntactic constraints. Experiments demonstrate considerably improved optimization over existing approaches across a broad range of constraints, including the popular setting where syntax is governed by a context-free grammar.

Model Interpretability through the lens of Computational Complexity

- Pablo Barceló³, Mikaël Monet, Jorge Pérez, Bernardo Subercaseaux
- abstract@[open-review](#): In spite of several claims stating that some models are more interpretable than others --e.g., "linear models are more interpretable than deep neural networks"-- we still lack a principled notion of interpretability that allows us to formally compare among different classes of models. We make a step towards such a theory by studying whether folklore interpretability claims have a correlate in terms of computational complexity theory. We focus on post-hoc explainability queries that, intuitively, attempt to answer why individual inputs are classified in a certain way by a given model. In a nutshell, we say that a class C1 of models is more interpretable than another class C2, if the computational complexity of answering post-hoc queries for models in C2 is higher than for C1. We prove that this notion provides a good theoretical counterpart to current beliefs on the interpretability of models; in particular, we show that under our definition and assuming standard complexity-theoretical assumptions (such as P!=NP), both linear and tree-based models are strictly more interpretable than neural networks. Our complexity analysis, however, does not provide a clear-cut difference between linear and tree-based models, as we obtain different results depending on the particular {post-hoc explanations} considered. Finally, by applying a finer complexity analysis based on parameterized complexity, we are able to prove a theoretical result suggesting that shallow neural networks are more interpretable than deeper ones.

Markovian Score Climbing: Variational Inference with KL(p||q)

- Christian Naesseth, Fredrik Lindsten, David Blei
- abstract@[open-review](#): Modern variational inference (VI) uses stochastic gradients to avoid intractable expectations, enabling large-scale probabilistic inference in complex models. VI posits a family of approximating distributions q and then finds the member of that family that is closest to the exact posterior p . Traditionally, VI algorithms minimize the "exclusive Kullback-Leibler (KL)" $= \text{KL}(q||p)$, often for computational convenience. Recent research, however, has also focused on the "inclusive KL" $= \text{KL}(p||q)$, which has good statistical properties that makes it more appropriate for certain inference problems. This paper develops a simple algorithm for reliably minimizing the inclusive KL using stochastic gradients with vanishing bias. This method, which we call Markovian score climbing (MSC), converges to a local optimum of the inclusive KL. It does not suffer from the systematic errors inherent in existing methods, such as Reweighted Wake-Sleep and Neural Adaptive Sequential Monte Carlo, which lead to bias in their final estimates. We illustrate convergence on a toy model and demonstrate the utility of MSC on Bayesian probit regression for classification as well as a stochastic volatility model for financial data.

Improved Analysis of Clipping Algorithms for Non-convex Optimization

- Bohang Zhang, Jikai Jin, Cong Fang, Liwei Wang
- abstract@[open-review](#): Gradient clipping is commonly used in training deep neural networks partly due to its practicability in relieving the exploding gradient problem. Recently, \citet{zhang2019gradient} show that clipped (stochastic) Gradient Descent (GD) converges faster than vanilla GD via introducing a new assumption called (L_0, L_1) -smoothness, which characterizes the violent fluctuation of gradients typically encountered in deep neural networks. However, their iteration complexities on the problem-dependent parameters are rather pessimistic, and theoretical justification of clipping combined with other crucial techniques, e.g. momentum acceleration, are still lacking. In this paper, we bridge the gap by presenting a general framework to study the clipping algorithms, which also takes momentum methods into consideration. We provide convergence analysis of the framework in both deterministic and stochastic setting, and demonstrate the tightness of our results by comparing them with existing lower bounds. Our results imply that the efficiency of clipping methods will not degenerate even in highly non-smooth regions of the landscape. Experiments confirm the superiority of clipping-based methods in deep learning tasks.

Bias no more: high-probability data-dependent regret bounds for adversarial bandits and MDPs

- Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang
- abstract@[open-review](#): We develop a new approach to obtaining high probability regret bounds for online learning with bandit feedback against an adaptive adversary. While existing approaches all require carefully constructing optimistic and biased loss estimators, our approach uses standard unbiased estimators and relies on a simple increasing learning rate schedule, together with the help of logarithmically homogeneous self-concordant barriers and a strengthened Freedman's inequality. Besides its simplicity, our approach enjoys several advantages. First, the obtained high-probability regret bounds are data-dependent and could be much smaller than the worst-case bounds, which resolves an open problem asked by Neu (2015). Second, resolving another open problem of Bartlett et al. (2008) and Abernethy and Rakhlin (2009), our approach leads to the first general and efficient algorithm with a high-probability regret bound for adversarial linear bandits, while previous methods are either inefficient or only applicable to specific action sets. Finally, our approach can also be applied to learning adversarial Markov Decision Processes and provides the first algorithm with a high-probability small-loss bound for this problem.

A Ranking-based, Balanced Loss Function Unifying Classification and Localisation in Object Detection

- Kemal Oksuz, Baris Can Cam, Emre Akbas, Sinan Kalkan
- abstract@[open-review](#): We propose average Localisation-Recall-Precision (aLRP), a unified, bounded, balanced and ranking-based loss function for both classification and localisation tasks in object detection. aLRP extends the Localisation-Recall-Precision (LRP) performance metric (Oksuz et al., 2018) inspired from how Average Precision (AP) Loss extends precision to a ranking-based loss function for classification (Chen et al., 2020). aLRP has the following distinct advantages: (i) aLRP is the first ranking-based loss function for both classification and localisation tasks. (ii) Thanks to using ranking for both tasks, aLRP naturally enforces high-quality localisation for high-precision classification. (iii) aLRP provides provable balance between positives and negatives. (iv) Compared to on average ~6 hyperparameters in the loss functions of state-of-the-art detectors, aLRP Loss has only one hyperparameter, which we did not tune in practice. On the COCO dataset, aLRP Loss improves its ranking-based predecessor, AP Loss, up to around 5 AP points, achieves 48.9 AP without test time augmentation and outperforms all one-stage detectors. Code available at: <https://github.com/kemaloksuz/aLRPLoss>.

StratLearner: Learning a Strategy for Misinformation Prevention in Social Networks

- Guangmo Tong
- abstract@[open-review](#): Given a combinatorial optimization problem taking an input, can we learn a strategy to solve it from the examples of input-solution pairs without knowing its objective function? In this paper, we consider such a setting and study the misinformation prevention problem. Given the examples of attacker-protector pairs, our goal is to learn a strategy to compute protectors against future attackers, without the need of knowing the underlying diffusion model. To this end, we design a structured prediction framework, where the main idea is to parameterize the scoring function using random features constructed through distance functions on randomly sampled subgraphs, which leads to a kernelized scoring function with weights learnable via the large margin method. Evidenced by experiments, our method can produce near-optimal protectors without using any information of the diffusion model, and it outperforms other possible graph-based and learning-based methods by an evident margin.

[A Unified Switching System Perspective and Convergence Analysis of Q-Learning Algorithms](#)

- Donghwan Lee, Niao He
- abstract@[open-review](#): This paper develops a novel and unified framework to analyze the convergence of a large family of Q-learning algorithms from the switching system perspective. We show that the nonlinear ODE models associated with Q-learning and many of its variants can be naturally formulated as affine switching systems. Building on their asymptotic stability, we obtain a number of interesting results: (i) we provide a simple ODE analysis for the convergence of asynchronous Q-learning under relatively weak assumptions; (ii) we establish the first convergence analysis of the averaging Q-learning algorithm; and (iii) we derive a new sufficient condition for the convergence of Q-learning with linear function approximation.

[Kernel Alignment Risk Estimator: Risk Prediction from Training Data](#)

- Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clement Hongler, Franck Gabriel
- abstract@[open-review](#): We study the risk (i.e. generalization error) of Kernel Ridge Regression (KRR) for a kernel K with ridge $\lambda > 0$ and i.i.d. observations. For this, we introduce two objects: the Signal Capture Threshold (SCT) and the Kernel Alignment Risk Estimator (KARE). The SCT $\vartheta_{K,\lambda}$ is a function of the data distribution: it can be used to identify the components of the data that the KRR predictor captures, and to approximate the (expected) KRR risk. This then leads to a KRR risk approximation by the KARE $\rho_{K,\lambda}$, an explicit function of the training data, agnostic of the true data distribution. We phrase the regression problem in a functional setting. The key results then follow from a finite-size adaptation of the resolvent method for general Wishart random matrices. Under a natural universality assumption (that the KRR moments depend asymptotically on the first two moments of the observations) we capture the mean and variance of the KRR predictor. We numerically investigate our findings on the Higgs and MNIST datasets for various classical kernels: the KARE gives an excellent approximation of the risk. This supports our universality hypothesis. Using the KARE, one can compare choices of Kernels and hyperparameters directly from the training set. The KARE thus provides a promising data-dependent procedure to select Kernels that generalize well.

[Calibrating CNNs for Lifelong Learning](#)

- Pravendra Singh, Vinay Kumar Verma, Pratik Mazumder, Lawrence Carin, Piyush Rai
- abstract@[open-review](#): We present an approach for lifelong/continual learning of convolutional neural networks (CNN) that does not suffer from the problem of catastrophic forgetting when moving from one task to the other. We show that the activation maps generated by the CNN trained on the old task can be calibrated using very few calibration parameters, to become relevant to the new task. Based on this, we calibrate the activation maps produced by each network layer using spatial and channel-wise calibration modules and train only these calibration parameters for each new task in order to perform lifelong learning. Our calibration modules introduce significantly less computation and parameters as compared to the approaches that dynamically expand the network. Our approach is immune to catastrophic forgetting since we store the task-adaptive calibration parameters, which contain all the task-specific knowledge and is exclusive to each task. Further, our approach does not require storing data samples from the old tasks, which is done by many replay based methods. We perform extensive experiments on multiple benchmark datasets (SVHN, CIFAR, ImageNet, and MS-Celeb), all of which show substantial improvements over state-of-the-art methods (e.g., a 29% absolute increase in accuracy on CIFAR-100 with 10 classes at a time). On large-scale datasets, our approach yields 23.8% and 9.7% absolute increase in accuracy on ImageNet-100 and MS-Celeb-10K datasets, respectively, by employing very few (0.51% and 0.35% of model parameters) task-adaptive calibration parameters.

[Online Convex Optimization Over Erdos-Renyi Random Networks](#)

- Jinlong Lei, Peng Yi, Yiguang Hong, Jie Chen, Guodong Shi
- abstract@[open-review](#): The work studies how node-to-node communications over an Erdős-Rényi random network influence distributed online convex optimization, which is vital in solving large-scale machine learning in antagonistic or changing environments. At per step, each node (computing unit) makes a local decision, experiences a loss evaluated with a convex function, and communicates the decision with other nodes over a network. The node-to-node communications are described by the Erdős-Rényi rule, where independently each link takes place with a probability p over a prescribed connected graph. The objective is to minimize the system-wide loss accumulated over a finite time horizon. We consider standard distributed gradient descents with full gradients, one-point bandits and two-points bandits for convex and strongly convex losses, respectively. We establish how the regret bounds scale with respect to time horizon T , network size N , decision dimension d , and an algebraic network connectivity. The regret bounds scaling with respect to T match those obtained by state-of-the-art algorithms and fundamental limits in the corresponding centralized online optimization problems, e.g., $\mathcal{O}(\sqrt{T})$ and $\mathcal{O}(\ln(T))$ regrets are established for convex and strongly convex losses with full gradient feedback and two-points information, respectively. For classical Erdős-Rényi networks over all-to-all possible node communications, the regret scalings with respect to the probability p are analytically established, based on which the tradeoff between the communication overhead and computation accuracy is clearly demonstrated. Numerical studies have validated the theoretical findings.

[Robustness of Bayesian Neural Networks to Gradient-Based Attacks](#)

- Ginevra Carbone, Matthew Wicker, Luca Laurenti, Andrea Patane', Luca Bortolussi, Guido Sanguinetti
- abstract@[open-review](#): Vulnerability to adversarial attacks is one of the principal hurdles to the adoption of deep learning in safety-critical applications. Despite significant efforts, both practical and theoretical, the problem remains open. In this paper, we analyse the geometry of adversarial attacks in the large-data, overparametrized limit for Bayesian Neural Networks (BNNs). We show that, in the limit, vulnerability to gradient-based attacks arises as a result of degeneracy in the data distribution, i.e., when the data lies on a lower-dimensional submanifold of the ambient space. As a direct consequence, we demonstrate that in the limit BNN posteriors are robust to gradient-based adversarial attacks. Experimental results on the MNIST and Fashion MNIST datasets with BNNs trained with Hamiltonian Monte Carlo and Variational Inference support this line of argument, showing that BNNs can display both high accuracy and robustness to gradient based adversarial attacks.

[Parametric Instance Classification for Unsupervised Visual Feature learning](#)

- Yue Cao, Zhenda Xie, Bin Liu, Yutong Lin, Zheng Zhang, Han Hu
- abstract@[open-review](#): This paper presents parametric instance classification (PIC) for unsupervised visual feature learning. Unlike the state-of-the-art approaches which do instance discrimination in a dual-branch non-parametric fashion, PIC directly performs a one-branch parametric instance classification, revealing a simple framework similar to supervised classification and without the need to address the information leakage issue. We show that the simple PIC framework can be as effective as the state-of-the-art approaches, i.e. SimCLR and MoCo v2, by adapting several common component settings used in the state-of-the-art approaches. We also propose two novel techniques to further improve effectiveness and practicality of PIC: 1) a sliding-window data scheduler, instead of the previous epoch-based data scheduler, which addresses the extremely infrequent instance visiting issue in PIC and improves the effectiveness; 2) a negative sampling and weight update correction approach to reduce the training time and GPU memory consumption, which also enables application of PIC to almost unlimited training images. We hope that the PIC framework can serve as a simple baseline to facilitate future study. The code and network configurations are available at <https://github.com/b10/PIC>.

[Sparse Weight Activation Training](#)

- Md Aamir Raihan, Tor Aamodt

- abstract@[open-review](#): Neural network training is computationally and memory intensive. Sparse training can reduce the burden on emerging hardware platforms designed to accelerate sparse computations, but it can also affect network convergence. In this work, we propose a novel CNN training algorithm called Sparse Weight Activation Training (SWAT). SWAT is more computation and memory-efficient than conventional training. SWAT modifies back-propagation based on the empirical insight that convergence during training tends to be robust to the elimination of (i) small magnitude weights during the forward pass and (ii) both small magnitude weights and activations during the backward pass. We evaluate SWAT on recent CNN architectures such as ResNet, VGG, DenseNet and WideResNet using CIFAR-10, CIFAR-100 and ImageNet datasets. For ResNet-50 on ImageNet SWAT reduces total floating-point operations (FLOPs) during training by 80% resulting in a 3.3x training speedup when run on a simulated sparse learning accelerator representative of emerging platforms while incurring only 1.63% reduction in validation accuracy. Moreover, SWAT reduces memory footprint during the backward pass by 23% to 50% for activations and 50% to 90% for weights. Code is available at <https://github.com/AamirRaihan/SWAT>.

[Collapsing Bandits and Their Application to Public Health Intervention](#)

- Aditya Mate, Jackson Killian, Haifeng Xu, Andrew Perrault, Milind Tambe
- abstract@[open-review](#): We propose and study Collapsing Bandits, a new restless multi-armed bandit (RMAB) setting in which each arm follows a binary-state Markovian process with a special structure: when an arm is played, the state is fully observed, thusâ€œcollapsingâ€ any uncertainty, but when an arm is passive, no observation is made, thus allowing uncertainty to evolve. The goal is to keep as many arms in the â€œgoodâ€ state as possible by planning a limited budget of actions per round. Such Collapsing Bandits are natural models for many healthcare domains in which health workers must simultaneously monitor patients and deliver interventions in a way that maximizes the health of their patient cohort. Our main contributions are as follows: (i) Building on the Whittle index technique for RMABs, we derive conditions under which the Collapsing Bandits problem is indexable. Our derivation hinges on novel conditions that characterize when the optimal policies may take the form of eitherâ€œforwardâ€ or â€œreverseâ€ threshold policies. (ii) We exploit the optimality of threshold policies to build fast algorithms for computing the Whittle index, including a closed-form. (iii) We evaluate our algorithm on several data distributions including data from a real-world healthcare task in which a worker must monitor and deliver interventions to maximize their patientsâ€™ adherence to tuberculosis medication. Our algorithm achieves a 3-order-of-magnitude speedup compared to state-of-the-art RMAB techniques, while achieving similar performance. The code is available at:https://github.com/AdityaMate/collapsing_bandits

[Neural Sparse Voxel Fields](#)

- Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, Christian Theobalt
- abstract@[open-review](#): Photo-realistic free-viewpoint rendering of real-world scenes using classical computer graphics techniques is challenging, because it requires the difficult step of capturing detailed appearance and geometry models. Recent studies have demonstrated promising results by learning scene representations that implicitly encodes both geometry and appearance without 3D supervision. However, existing approaches in practice often show blurry renderings caused by the limited network capacity or the difficulty in finding accurate intersections of camera rays with the scene geometry. Synthesizing high-resolution imagery from these representations often requires time-consuming optical ray marching. In this work, we introduce Neural Sparse Voxel Fields (NSVF), a new neural scene representation for fast and high-quality free-viewpoint rendering. The NSVF defines a series of voxel-bounded implicit fields organized in a sparse voxel octree to model local properties in each cell. We progressively learn the underlying voxel structures with a differentiable ray-marching operation from only a set of posed RGB images. With the sparse voxel octree structure, rendering novel views at inference time can be accelerated by skipping the voxels without relevant scene content. Our method is over 10 times faster than the state-of-the-art while achieving higher quality results. Furthermore, by utilizing an explicit sparse voxel representation, our method can be easily applied to scene editing and scene composition. we also demonstrate various kinds of challenging tasks, including multi-object learning, free-viewpoint rendering of a moving human, and large-scale scene rendering.

[A Flexible Framework for Designing Trainable Priors with Adaptive Smoothing and Game Encoding](#)

- Bruno Lecouat, Jean Ponce, Julien Mairal
- abstract@[open-review](#): We introduce a general framework for designing and training neural network layers whose forward passes can be interpreted as solving non-smooth convex optimization problems, and whose architectures are derived from an optimization algorithm. We focus on convex games, solved by local agents represented by the nodes of a graph and interacting through regularization functions. This approach is appealing for solving imaging problems, as it allows the use of classical image priors within deep models that are trainable end to end. The priors used in this presentation include variants of total variation, Laplacian regularization, bilateral filtering, sparse coding on learned dictionaries, and non-local self similarities. Our models are fully interpretable as well as parameter and data efficient. Our experiments demonstrate their effectiveness on a large diversity of tasks ranging from image denoising and compressed sensing for fMRI to dense stereo matching.

[The Discrete Gaussian for Differential Privacy](#)

- ClÃ©ment L. Canonne, Gautam Kamath, Thomas Steinke
- abstract@[open-review](#): A key tool for building differentially private systems is adding Gaussian noise to the output of a function evaluated on a sensitive dataset. Unfortunately, using a continuous distribution presents several practical challenges. First and foremost, finite computers cannot exactly represent samples from continuous distributions, and previous work has demonstrated that seemingly innocuous numerical errors can entirely destroy privacy. Moreover, when the underlying data is itself discrete (e.g., population counts), adding continuous noise makes the result less interpretable. With these shortcomings in mind, we introduce and analyze the discrete Gaussian in the context of differential privacy. Specifically, we theoretically and experimentally show that adding discrete Gaussian noise provides essentially the same privacy and accuracy guarantees as the addition of continuous Gaussian noise. We also present an simple and efficient algorithm for exact sampling from this distribution. This demonstrates its applicability for privately answering counting queries, or more generally, low-sensitivity integer-valued queries.

[Robust Sub-Gaussian Principal Component Analysis and Width-Independent Schatten Packing](#)

- Arun Jambulapati, Jerry Li, Kevin Tian
- abstract@[open-review](#): We develop two methods for the following fundamental statistical task: given an $\$\\epsilon\$$ -corrupted set of n samples from a d -dimensional sub-Gaussian distribution, return an approximate top eigenvector of the covariance matrix. Our first robust PCA algorithm runs in polynomial time, returns a $1 - O(\\epsilon \\log \\epsilon^{-1})$ -approximate top eigenvector, and is based on a simple iterative filtering approach. Our second, which attains a slightly worse approximation factor, runs in nearly-linear time and sample complexity under a mild spectral gap assumption. These are the first polynomial-time algorithms yielding non-trivial information about the covariance of a corrupted sub-Gaussian distribution without requiring additional algebraic structure of moments. As a key technical tool, we develop the first width-independent solvers for Schatten- p norm packing semidefinite programs, giving a $(1 + \\epsilon)$ -approximate solution in $O(p \\log(t \\frac{n}{\\epsilon}))$ input-sparsity time iterations (where n, d are problem dimensions).

[Adaptive Importance Sampling for Finite-Sum Optimization and Sampling with Decreasing Step-Sizes](#)

- Ayoub El Hanchi, David Stephens
- abstract@[open-review](#): Reducing the variance of the gradient estimator is known to improve the convergence rate of stochastic gradient-based optimization and sampling algorithms. One way of achieving variance reduction is to design importance sampling strategies. Recently, the problem of

designing such schemes was formulated as an online learning problem with bandit feedback, and algorithms with sub-linear static regret were designed. In this work, we build on this framework and propose a simple and efficient algorithm for adaptive importance sampling for finite-sum optimization and sampling with decreasing step-sizes. Under standard technical conditions, we show that our proposed algorithm achieves $O(T^{1/2})$ and $O(T^{1/3})$ dynamic regret for SGD and SGLD respectively when run with $O(1/t)$ step sizes. We achieve this dynamic regret bound by leveraging our knowledge of the dynamics defined by the algorithm, and combining ideas from online learning and variance-reduced stochastic optimization. We validate empirically the performance of our algorithm and identify settings in which it leads to significant improvements.

[Learning efficient task-dependent representations with synaptic plasticity](#)

- Colin Bredenberg, Eero Simoncelli, Cristina Savin
- abstract@[open-review](#): Neural populations encode the sensory world imperfectly: their capacity is limited by the number of neurons, availability of metabolic and other biophysical resources, and intrinsic noise. The brain is presumably shaped by these limitations, improving efficiency by discarding some aspects of incoming sensory streams, while preferentially preserving commonly occurring, behaviorally-relevant information. Here we construct a stochastic recurrent neural circuit model that can learn efficient, task-specific sensory codes using a novel form of reward-modulated Hebbian synaptic plasticity. We illustrate the flexibility of the model by training an initially unstructured neural network to solve two different tasks: stimulus estimation, and stimulus discrimination. The network achieves high performance in both tasks by appropriately allocating resources and using its recurrent circuitry to best compensate for different levels of noise. We also show how the interaction between stimulus priors and task structure dictates the emergent network representations.

[A Contour Stochastic Gradient Langevin Dynamics Algorithm for Simulations of Multi-modal Distributions](#)

- Wei Deng, Guang Lin, Faming Liang
- abstract@[open-review](#): We propose an adaptively weighted stochastic gradient Langevin dynamics algorithm (SGLD), so-called contour stochastic gradient Langevin dynamics (CSGLD), for Bayesian learning in big data statistics. The proposed algorithm is essentially a scalable dynamic importance sampler, which automatically flattens the target distribution such that the simulation for a multi-modal distribution can be greatly facilitated. Theoretically, we prove a stability condition and establish the asymptotic convergence of the self-adapting parameter to a unique fixed-point, regardless of the non-convexity of the original energy function; we also present an error analysis for the weighted averaging estimators. Empirically, the CSGLD algorithm is tested on multiple benchmark datasets including CIFAR10 and CIFAR100. The numerical results indicate its superiority over the existing state-of-the-art algorithms in training deep neural networks.

[Error Bounds of Imitating Policies and Environments](#)

- Tian Xu, Ziniu Li, Yang Yu
- abstract@[open-review](#): Imitation learning trains a policy by mimicking expert demonstrations. Various imitation methods were proposed and empirically evaluated, meanwhile, their theoretical understanding needs further studies. In this paper, we firstly analyze the value gap between the expert policy and imitated policies by two imitation methods, behavioral cloning and generative adversarial imitation. The results support that generative adversarial imitation can reduce the compounding errors compared to behavioral cloning, and thus has a better sample complexity. Noticed that by considering the environment transition model as a dual agent, imitation learning can also be used to learn the environment model. Therefore, based on the bounds of imitating policies, we further analyze the performance of imitating environments. The results show that environment models can be more effectively imitated by generative adversarial imitation than behavioral cloning, suggesting a novel application of adversarial imitation for model-based reinforcement learning. We hope these results could inspire future advances in imitation learning and model-based reinforcement learning.

[Disentangling Human Error from Ground Truth in Segmentation of Medical Images](#)

- Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Cicarrelli, Frederik Barkhof, Daniel Alexander
- abstract@[open-review](#): Recent years have seen increasing use of supervised learning methods for segmentation tasks. However, the predictive performance of these algorithms depends on the quality of labels. This problem is particularly pertinent in the medical image domain, where both the annotation cost and inter-observer variability are high. In a typical label acquisition process, different human experts provide their estimates of the ``true'' segmentation labels under the influence of their own biases and competence levels. Treating these noisy labels blindly as the ground truth limits the performance that automatic segmentation algorithms can achieve. In this work, we present a method for jointly learning, from purely noisy observations alone, the reliability of individual annotators and the true segmentation label distributions, using two coupled CNNs. The separation of the two is achieved by encouraging the estimated annotators to be maximally unreliable while achieving high fidelity with the noisy training data. We first define a toy segmentation dataset based on MNIST and study the properties of the proposed algorithm. We then demonstrate the utility of the method on three public medical imaging segmentation datasets with simulated (when necessary) and real diverse annotations: 1) MSLSC (multiple-sclerosis lesions); 2) BraTS (brain tumours); 3) LIDC-IDRI (lung abnormalities). In all cases, our method outperforms competing methods and relevant baselines particularly in cases where the number of annotations is small and the amount of disagreement is large. The experiments also show strong ability to capture the complex spatial characteristics of annotators' mistakes. Our code is available at \url{https://github.com/moucheng2017/LearnNoisyLabelsMedicalImages}.

[Consequences of Misaligned AI](#)

- Simon Zhuang, Dylan Hadfield-Menell
- abstract@[open-review](#): AI systems often rely on two key components: a specified goal or reward function and an optimization algorithm to compute the optimal behavior for that goal. This approach is intended to provide value for a principal: the user on whose behalf the agent acts. The objectives given to these agents often refer to a partial specification of the principal's goals. We consider the cost of this incompleteness by analyzing a model of a principal and an agent in a resource constrained world where the L features of the state correspond to different sources of utility for the principal. We assume that the reward function given to the agent only has support on $J < L$ features. The contributions of our paper are as follows: 1) we propose a novel model of an incomplete principal-agent problem from artificial intelligence; 2) we provide necessary and sufficient conditions under which indefinitely optimizing for any incomplete proxy objective leads to arbitrarily low overall utility; and 3) we show how modifying the setup to allow reward functions that reference the full state or allowing the principal to update the proxy objective over time can lead to higher utility solutions. The results in this paper argue that we should view the design of reward functions as an interactive and dynamic process and identifies a theoretical scenario where some degree of interactivity is desirable.

[Promoting Coordination through Policy Regularization in Multi-Agent Deep Reinforcement Learning](#)

- Julien Roy, Paul Barde, Félix Harvey, Derek Nowrouzezahrai, Chris Pal
- abstract@[open-review](#): In multi-agent reinforcement learning, discovering successful collective behaviors is challenging as it requires exploring a joint action space that grows exponentially with the number of agents. While the tractability of independent agent-wise exploration is appealing, this approach fails on tasks that require elaborate group strategies. We argue that coordinating the agents' policies can guide their exploration and we investigate techniques to promote such an inductive bias. We propose two policy regularization methods: TeamReg, which is based on inter-agent action predictability and CoachReg that relies on synchronized behavior selection. We evaluate each approach on four challenging continuous control tasks with sparse rewards that require varying levels of coordination as well as on the discrete action Google Research Football environment. Our experiments show

improved performance across many cooperative multi-agent problems. Finally, we analyze the effects of our proposed methods on the policies that our agents learn and show that our methods successfully enforce the qualities that we propose as proxies for coordinated behaviors.

[Emergent Reciprocity and Team Formation from Randomized Uncertain Social Preferences](#)

- Bowen Baker
- abstract@[open-review](#): Multi-agent reinforcement learning (MARL) has shown recent success in increasingly complex fixed-team zero-sum environments. However, the real world is not zero-sum nor does it have fixed teams; humans face numerous social dilemmas and must learn when to cooperate and when to compete. To successfully deploy agents into the human world, it may be important that they be able to understand and help in our conflicts. Unfortunately, selfish MARL agents typically fail when faced with social dilemmas. In this work, we show evidence of emergent direct reciprocity, indirect reciprocity and reputation, and team formation when training agents with randomized uncertain social preferences (RUSP), a novel environment augmentation that expands the distribution of environments agents play in. RUSP is generic and scalable; it can be applied to any multi-agent environment without changing the original underlying game dynamics or objectives. In particular, we show that with RUSP these behaviors can emerge and lead to higher social welfare equilibria in both classic abstract social dilemmas like Iterated Prisoner's Dilemma as well in more complex intertemporal environments.

[Hitting the High Notes: Subset Selection for Maximizing Expected Order Statistics](#)

- Aranyak Mehta, Uri Nadav, Alexandros Psomas, Aviad Rubinstein
- abstract@[open-review](#): We consider the fundamental problem of selecting k out of n random variables in a way that the expected highest or second-highest value is maximized. This question captures several applications where we have uncertainty about the quality of candidates (e.g. auction bids, search results) and have the capacity to explore only a small subset due to an exogenous constraint. For example, consider a second price auction where system constraints (e.g., costly retrieval or model computation) allow the participation of only k out of n bidders, and the goal is to optimize the expected efficiency (highest bid) or expected revenue (second highest bid).

We study the case where we are given an explicit description of each random variable. We give a PTAS for the problem of maximizing the expected highest value. For the second-highest value, we prove a hardness result: assuming the Planted Clique Hypothesis, there is no constant factor approximation algorithm that runs in polynomial time. Surprisingly, under the assumption that each random variable has monotone hazard rate (MHR), a simple score-based algorithm, namely picking the k random variables with the largest \sqrt{k} top quantile value, is a constant approximation to the expected highest and second highest value, simultaneously.

[Towards Scale-Invariant Graph-related Problem Solving by Iterative Homogeneous GNNs](#)

- Hao Tang, Zhaoyang Huang, Jiayuan Gu, Bao-Liang Lu, Hao Su
- abstract@[open-review](#): Current graph neural networks (GNNs) lack generalizability with respect to scales (graph sizes, graph diameters, edge weights, etc.) when solving many graph analysis problems. Taking the perspective of synthesizing graph theory programs, we propose several extensions to address the issue. First, inspired by the dependency of iteration number of common graph theory algorithms on graph size, we learn to terminate the message passing process in GNNs adaptively according to the computation progress. Second, inspired by the fact that many graph theory algorithms are homogeneous with respect to graph weights, we introduce homogeneous transformation layers that are universal homogeneous function approximators, to convert ordinary GNNs to be homogeneous. Experimentally, we show that our GNN can be trained from small-scale graphs but generalize well to large-scale graphs for a number of basic graph theory problems. It also shows generalizability for applications of multi-body physical simulation and image-based navigation problems.

[Regret Bounds without Lipschitz Continuity: Online Learning with Relative-Lipschitz Losses](#)

- Yihan Zhou, Victor Sanches Portella, Mark Schmidt, Nicholas Harvey
- abstract@[open-review](#): In online convex optimization (OCO), Lipschitz continuity of the functions is commonly assumed in order to obtain sublinear regret. Moreover, many algorithms have only logarithmic regret when these functions are also strongly convex. Recently, researchers from convex optimization proposed the notions of relative Lipschitz continuity" and relative strong convexity". Both of the notions are generalizations of their classical counterparts. It has been shown that subgradient methods in the relative setting have performance analogous to their performance in the classical setting. In this work, we consider OCO for relative Lipschitz and relative strongly convex functions. We extend the known regret bounds for classical OCO algorithms to the relative setting. Specifically, we show regret bounds for the follow the regularized leader algorithms and a variant of online mirror descent. Due to the generality of these methods, these results yield regret bounds for a wide variety of OCO algorithms. Furthermore, we further extend the results to algorithms with extra regularization such as regularized dual averaging.

[The Lottery Ticket Hypothesis for Pre-trained BERT Networks](#)

- Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, Michael Carbin
- abstract@[open-review](#): In natural language processing (NLP), enormous pre-trained models like BERT have become the standard starting point for training on a range of downstream tasks, and similar trends are emerging in other areas of deep learning. In parallel, work on the lottery ticket hypothesis has shown that models for NLP and computer vision contain smaller matching subnetworks capable of training in isolation to full accuracy and transferring to other tasks. In this work, we combine these observations to assess whether such trainable, transferrable subnetworks exist in pre-trained BERT models. For a range of downstream tasks, we indeed find matching subnetworks at 40% to 90% sparsity. We find these subnetworks at (pre-trained) initialization, a deviation from prior NLP research where they emerge only after some amount of training. Subnetworks found on the masked language modeling task (the same task used to pre-train the model) transfer universally; those found on other tasks transfer in a limited fashion if at all. As large-scale pre-training becomes an increasingly central paradigm in deep learning, our results demonstrate that the main lottery ticket observations remain relevant in this context. Codes available at <https://github.com/VITA-Group/BERT-Tickets>.

[Label-Aware Neural Tangent Kernel: Toward Better Generalization and Local Elasticity](#)

- Shuxiao Chen, Hangfeng He, Weijie Su
- abstract@[open-review](#): As a popular approach to modeling the dynamics of training overparametrized neural networks (NNs), the neural tangent kernels (NTK) are known to fall behind real-world NNs in generalization ability. This performance gap is in part due to the "label agnostic" nature of the NTK, which renders the resulting kernel not as "locally elastic" as $NNs \sim \text{citet{he2019local}}$. In this paper, we introduce a novel approach from the perspective of "label-awareness" to reduce this gap for the NTK. Specifically, we propose two label-aware kernels that are each a superimposition of a label-agnostic part and a hierarchy of label-aware parts with increasing complexity of label dependence, using the Hoeffding decomposition. Through both theoretical and empirical evidence, we show that the models trained with the proposed kernels better simulate NNs in terms of generalization ability and local elasticity.

[Beyond Perturbations: Learning Guarantees with Arbitrary Adversarial Test Examples](#)

- Shafi Goldwasser, Adam Tauman Kalai, Yael Kalai, Omar Montasser

- abstract@[open-review](#): We present a transductive learning algorithm that takes as input training examples from a distribution P and arbitrary (unlabeled) test examples, possibly chosen by an adversary. This is unlike prior work that assumes that test examples are small perturbations of P. Our algorithm outputs a selective classifier, which abstains from predicting on some examples. By considering selective transductive learning, we give the first nontrivial guarantees for learning classes of bounded VC dimension with arbitrary train and test distributions—no prior guarantees were known even for simple classes of functions such as intervals on the line. In particular, for any function in a class C of bounded VC dimension, we guarantee a low test error rate and a low rejection rate with respect to P. Our algorithm is efficient given an Empirical Risk Minimizer (ERM) for C. Our guarantees hold even for test examples chosen by an unbounded white-box adversary. We also give guarantees for generalization, agnostic, and unsupervised settings.

[AdvFlow: Inconspicuous Black-box Adversarial Attacks using Normalizing Flows](#)

- Hadi Mohaghegh Dolatabadi, Sarah Erfani, Christopher Leckie
- abstract@[open-review](#): Deep learning classifiers are susceptible to well-crafted, imperceptible variations of their inputs, known as adversarial attacks. In this regard, the study of powerful attack models sheds light on the sources of vulnerability in these classifiers, hopefully leading to more robust ones. In this paper, we introduce AdvFlow: a novel black-box adversarial attack method on image classifiers that exploits the power of normalizing flows to model the density of adversarial examples around a given target image. We see that the proposed method generates adversaries that closely follow the clean data distribution, a property which makes their detection less likely. Also, our experimental results show competitive performance of the proposed approach with some of the existing attack methods on defended classifiers.

[Few-shot Image Generation with Elastic Weight Consolidation](#)

- Yijun Li, Richard Zhang, Jingwan (Cynthia) Lu, Eli Shechtman
- abstract@[open-review](#): Few-shot image generation seeks to generate more data of a given domain, with only few available training examples. As it is unreasonable to expect to fully infer the distribution from just a few observations (e.g., emojis), we seek to leverage a large, related source domain as pretraining (e.g., human faces). Thus, we wish to preserve the diversity of the source domain, while adapting to the appearance of the target. We adapt a pretrained model, without introducing any additional parameters, to the few examples of the target domain. Crucially, we regularize the changes of the weights during this adaptation, in order to best preserve the information of the source dataset, while fitting the target. We demonstrate the effectiveness of our algorithm by generating high-quality results of different target domains, including those with extremely few examples (e.g., 10). We also analyze the performance of our method with respect to some important factors, such as the number of examples and the similarity between the source and target domain.

[On the Expressiveness of Approximate Inference in Bayesian Neural Networks](#)

- Andrew Foong, David Burt, Yingzhen Li, Richard Turner
- abstract@[open-review](#): While Bayesian neural networks (BNNs) hold the promise of being flexible, well-calibrated statistical models, inference often requires approximations whose consequences are poorly understood. We study the quality of common variational methods in approximating the Bayesian predictive distribution. For single-hidden layer ReLU BNNs, we prove a fundamental limitation in function-space of two of the most commonly used distributions defined in weight-space: mean-field Gaussian and Monte Carlo dropout. We find there are simple cases where neither method can have substantially increased uncertainty in between well-separated regions of low uncertainty. We provide strong empirical evidence that exact inference does not have this pathology, hence it is due to the approximation and not the model. In contrast, for deep networks, we prove a universality result showing that there exist approximate posteriors in the above classes which provide flexible uncertainty estimates. However, we find empirically that pathologies of a similar form as in the single-hidden layer case can persist when performing variational inference in deeper networks. Our results motivate careful consideration of the implications of approximate inference methods in BNNs.

[Non-Crossing Quantile Regression for Distributional Reinforcement Learning](#)

- Fan Zhou, Jianing Wang, Xingdong Feng
- abstract@[open-review](#): Distributional reinforcement learning (DRL) estimates the distribution over future returns instead of the mean to more efficiently capture the intrinsic uncertainty of MDPs. However, batch-based DRL algorithms cannot guarantee the non-decreasing property of learned quantile curves especially at the early training stage, leading to abnormal distribution estimates and reduced model interpretability. To address these issues, we introduce a general DRL framework by using non-crossing quantile regression to ensure the monotonicity constraint within each sampled batch, which can be incorporated with any well-known DRL algorithm. We demonstrate the validity of our method from both the theory and model implementation perspectives. Experiments on Atari 2600 Games show that some state-of-art DRL algorithms with the non-crossing modification can significantly outperform their baselines in terms of faster convergence speeds and better testing performance. In particular, our method can effectively recover the distribution information and thus dramatically increase the exploration efficiency when the reward space is extremely sparse.

[Dark Experience for General Continual Learning: a Strong, Simple Baseline](#)

- Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, SIMONE CALDERARA
- abstract@[open-review](#): Continual Learning has inspired a plethora of approaches and evaluation settings; however, the majority of them overlooks the properties of a practical scenario, where the data stream cannot be shaped as a sequence of tasks and offline training is not viable. We work towards General Continual Learning (GCL), where task boundaries blur and the domain and class distributions shift either gradually or suddenly. We address it through mixing rehearsal with knowledge distillation and regularization; our simple baseline, Dark Experience Replay, matches the network's logits sampled throughout the optimization trajectory, thus promoting consistency with its past. By conducting an extensive analysis on both standard benchmarks and a novel GCL evaluation setting (MNIST-360), we show that such a seemingly simple baseline outperforms consolidated approaches and leverages limited resources. We further explore the generalization capabilities of our objective, showing its regularization being beneficial beyond mere performance.

[Learning to Utilize Shaping Rewards: A New Approach of Reward Shaping](#)

- Yujing Hu, Weixun Wang, Hangtian Jia, Yixiang Wang, Yingfeng Chen, Jianye Hao, Feng Wu, Changjie Fan
- abstract@[open-review](#): Reward shaping is an effective technique for incorporating domain knowledge into reinforcement learning (RL). Existing approaches such as potential-based reward shaping normally make full use of a given shaping reward function. However, since the transformation of human knowledge into numeric reward values is often imperfect due to reasons such as human cognitive bias, completely utilizing the shaping reward function may fail to improve the performance of RL algorithms. In this paper, we consider the problem of adaptively utilizing a given shaping reward function. We formulate the utilization of shaping rewards as a bi-level optimization problem, where the lower level is to optimize policy using the shaping rewards and the upper level is to optimize a parameterized shaping weight function for true reward maximization. We formally derive the gradient of the expected true reward with respect to the shaping weight function parameters and accordingly propose three learning algorithms based on different assumptions. Experiments in sparse-reward cartpole and MuJoCo environments show that our algorithms can fully exploit beneficial shaping rewards, and meanwhile ignore unbeneficial shaping rewards or even transform them into beneficial ones.

[Neural encoding with visual attention](#)

- Meenakshi Khosla, Gia Ngo, Keith Jamison, Amy Kuceyeski, Mert Sabuncu
- abstract@[open-review](#): Visual perception is critically influenced by the focus of attention. Due to limited resources, it is well known that neural representations are biased in favor of attended locations. Using concurrent eye-tracking and functional Magnetic Resonance Imaging (fMRI) recordings from a large cohort of human subjects watching movies, we first demonstrate that leveraging gaze information, in the form of attentional masking, can significantly improve brain response prediction accuracy in a neural encoding model. Next, we propose a novel approach to neural encoding by including a trainable soft-attention module. Using our new approach, we demonstrate that it is possible to learn visual attention policies by end-to-end learning merely on fMRI response data, and without relying on any eye-tracking. Interestingly, we find that attention locations estimated by the model on independent data agree well with the corresponding eye fixation patterns, despite no explicit supervision to do so. Together, these findings suggest that attention modules can be instrumental in neural encoding models of visual stimuli.

[On the linearity of large non-linear models: when and why the tangent kernel is constant](#)

- Chaoyue Liu, Libin Zhu, Misha Belkin
- abstract@[open-review](#): The goal of this work is to shed light on the remarkable phenomenon of "transition to linearity" of certain neural networks as their width approaches infinity. We show that the "transition to linearity" of the model and, equivalently, constancy of the (neural) tangent kernel (NTK) result from the scaling properties of the norm of the Hessian matrix of the network as a function of the network width. We present a general framework for understanding the constancy of the tangent kernel via Hessian scaling applicable to the standard classes of neural networks. Our analysis provides a new perspective on the phenomenon of constant tangent kernel, which is different from the widely accepted "lazy training". Furthermore, we show that the "transition to linearity" is not a general property of wide neural networks and does not hold when the last layer of the network is non-linear. It is also not necessary for successful optimization by gradient descent.

[PLlay: Efficient Topological Layer based on Persistent Landscapes](#)

- Kwangho Kim, Jisu Kim, Manzil Zaheer, Joon Kim, Frederic Chazal, Larry Wasserman
- abstract@[open-review](#): We propose PLlay, a novel topological layer for general deep learning models based on persistence landscapes, in which we can efficiently exploit the underlying topological features of the input data structure. In this work, we show differentiability with respect to layer inputs, for a general persistent homology with arbitrary filtration. Thus, our proposed layer can be placed anywhere in the network and feed critical information on the topological features of input data into subsequent layers to improve the learnability of the networks toward a given task. A task-optimal structure of PLlay is learned during training via backpropagation, without requiring any input featurization or data preprocessing. We provide a novel adaptation for the DTM function-based filtration, and show that the proposed layer is robust against noise and outliers through a stability analysis. We demonstrate the effectiveness of our approach by classification experiments on various datasets.

[Decentralized Langevin Dynamics for Bayesian Learning](#)

- Anjaly Parayil, He Bai, Jemin George, Prudhvi Gurram
- abstract@[open-review](#): Motivated by decentralized approaches to machine learning, we propose a collaborative Bayesian learning algorithm taking the form of decentralized Langevin dynamics in a non-convex setting. Our analysis show that the initial KL-divergence between the Markov Chain and the target posterior distribution is exponentially decreasing while the error contributions to the overall KL-divergence from the additive noise is decreasing in polynomial time. We further show that the polynomial-term experiences speed-up with number of agents and provide sufficient conditions on the time-varying step-sizes to guarantee convergence to the desired distribution. The performance of the proposed algorithm is evaluated on a wide variety of machine learning tasks. The empirical results show that the performance of individual agents with locally available data is on par with the centralized setting with considerable improvement in the convergence rate.

[Shared Space Transfer Learning for analyzing multi-site fMRI data](#)

- Tony Muhammad Yousefnezhad, Alessandro Selvitella, Daoqiang Zhang, Andrew Greenshaw, Russell Greiner
- abstract@[open-review](#): Multi-voxel pattern analysis (MVPA) learns predictive models from task-based functional magnetic resonance imaging (fMRI) data, for distinguishing when subjects are performing different cognitive tasks e.g., watching movies or making decisions. MVPA works best with a well-designed feature set and an adequate sample size. However, most fMRI datasets are noisy, high-dimensional, expensive to collect, and with small sample sizes. Further, training a robust, generalized predictive model that can analyze homogeneous cognitive tasks provided by multi-site fMRI datasets has additional challenges. This paper proposes the Shared Space Transfer Learning (SSTL) as a novel transfer learning (TL) approach that can functionally align homogeneous multi-site fMRI datasets, and so improve the prediction performance in every site. SSTL first extracts a set of common features for all subjects in each site. It then uses TL to map these site-specific features to a site-independent shared space in order to improve the performance of the MVPA. SSTL uses a scalable optimization procedure that works effectively for high-dimensional fMRI datasets. The optimization procedure extracts the common features for each site by using a single-iteration algorithm and maps these site-specific common features to the site-independent shared space. We evaluate the effectiveness of the proposed method for transferring between various cognitive tasks. Our comprehensive experiments validate that SSTL achieves superior performance to other state-of-the-art analysis techniques.

[The Diversified Ensemble Neural Network](#)

- Shaofeng Zhang, Meng Liu, Junchi Yan
- abstract@[open-review](#): Ensemble is a general way of improving the accuracy and stability of learning models, especially for the generalization ability on small datasets. Compared with tree-based methods, relatively less works have been devoted to an in-depth study on effective ensemble design for neural networks. In this paper, we propose a principled ensemble technique by constructing the so-called diversified ensemble layer to combine multiple networks as individual modules. We theoretically show that each individual model in our ensemble layer corresponds to weights in the ensemble layer optimized in different directions. Meanwhile, the devised ensemble layer can be readily integrated into popular neural architectures, including CNNs, RNNs, and GCNs. Extensive experiments are conducted on public tabular datasets, images, and texts. By adopting weight sharing approach, the results show our method can notably improve the accuracy and stability of the original neural networks with ignorable extra time and space overhead.

[Inductive Quantum Embedding](#)

- Santosh Kumar Srivastava, Dinesh Khandelwal, Dhiraj Madan, Dinesh Garg, Hima Karanam, L Venkata Subramaniam
- abstract@[open-review](#): Quantum logic inspired embedding (aka Quantum Embedding (QE)) of a Knowledge-Base (KB) was proposed recently by Garg:2019. It is claimed that the QE preserves the logical structure of the input KB given in the form of unary and binary predicates hierarchy. Such structure preservation allows one to perform Boolean logic style deductive reasoning directly over these embedding vectors. The original QE idea, however, is limited to the transductive (not inductive) setting. Moreover, the original QE scheme runs quite slow on real applications involving millions of entities. This paper alleviates both of these key limitations. We start by reformulating the original QE problem to allow for the induction. On the way, we also underscore some interesting analytic and geometric properties of the solution and leverage them to design a faster training scheme. As an application, we show that one can achieve state-of-the-art performance on the well-known NLP task of fine-grained entity type classification by using the inductive QE approach. Our training runs 9-times faster than the original QE scheme on this task.

Variational Bayesian Unlearning

- Quoc Phong Nguyen, Bryan Kian Hsiang Low, Patrick Jaillet
- abstract@[open-review](#): This paper studies the problem of approximately unlearning a Bayesian model from a small subset of the training data to be erased. We frame this problem as one of minimizing the Kullback-Leibler divergence between the approximate posterior belief of model parameters after directly unlearning from erased data vs. the exact posterior belief from retraining with remaining data. Using the variational inference (VI) framework, we show that it is equivalent to minimizing an evidence upper bound which trades off between fully unlearning from erased data vs. not entirely forgetting the posterior belief given the full data (i.e., including the remaining data); the latter prevents catastrophic unlearning that can render the model useless. In model training with VI, only an approximate (instead of exact) posterior belief given the full data can be obtained, which makes unlearning even more challenging. We propose two novel tricks to tackle this challenge. We empirically demonstrate our unlearning methods on Bayesian models such as sparse Gaussian process and logistic regression using synthetic and real-world datasets.

Batched Coarse Ranking in Multi-Armed Bandits

- Nikolai Karpov, Qin Zhang
- abstract@[open-review](#): We study the problem of coarse ranking in the multi-armed bandits (MAB) setting, where we have a set of arms each of which is associated with an unknown distribution. The task is to partition the arms into clusters of predefined sizes, such that the mean of any arm in the i -th cluster is larger than that of any arm in the j -th cluster for any $j > i$. Coarse ranking generalizes a number of basic problems in MAB (e.g., best arm identification) and has many real-world applications. We initiate the study of the problem in the batched model where we can only have a small number of policy changes. We study both the fixed budget and fixed confidence variants in MAB, and propose algorithms and prove impossibility results which together give almost tight tradeoffs between the total number of arms pulls and the number of policy changes. We have tested our algorithms in both real and synthetic data; our experimental results have demonstrated the efficiency of the proposed methods.

Understanding and Improving Fast Adversarial Training

- Maksym Andriushchenko, Nicolas Flammarion
- abstract@[open-review](#): A recent line of work focused on making adversarial training computationally efficient for deep learning models. In particular, Wong et al. (2020) showed that ℓ_∞ -adversarial training with fast gradient sign method (FGSM) can fail due to a phenomenon called catastrophic overfitting, when the model quickly loses its robustness over a single epoch of training. We show that adding a random step to FGSM, as proposed in Wong et al. (2020), does not prevent catastrophic overfitting, and that randomness is not important per se --- its main role being simply to reduce the magnitude of the perturbation. Moreover, we show that catastrophic overfitting is not inherent to deep and overparametrized networks, but can occur in a single-layer convolutional network with a few filters. In an extreme case, even a single filter can make the network highly non-linear locally, which is the main reason why FGSM training fails. Based on this observation, we propose a new regularization method, GradAlign, that prevents catastrophic overfitting by explicitly maximizing the gradient alignment inside the perturbation set and improves the quality of the FGSM solution. As a result, GradAlign allows to successfully apply FGSM training also for larger ℓ_∞ -perturbations and reduce the gap to multi-step adversarial training. The code of our experiments is available at <https://github.com/tml-epfl/understanding-fast-adv-training>.

Coded Sequential Matrix Multiplication For Straggler Mitigation

- Nikhil Krishnan Muralee Krishnan, Seyederfan Hosseini, Ashish Khisti
- abstract@[open-review](#): In this work, we consider a sequence of J matrix multiplication jobs which needs to be distributed by a master across multiple worker nodes. For $i \in \{1, 2, \dots, J\}$, job- i begins in round- i and has to be completed by round- $(i+T)$. Previous works consider only the special case of $T=0$ and focus on coding across workers. We propose here two schemes with $T>0$, which feature coding across workers as well as the dimension of time. Our first scheme is a modification of the polynomial coding scheme introduced by Yu et al. and places no assumptions on the straggler model. Exploitation of the temporal dimension helps the scheme handle a larger set of straggler patterns than the polynomial coding scheme, for a given computational load per worker per round. The second scheme assumes a particular straggler model to further improve performance (in terms of encoding/decoding complexity). We develop theoretical results establishing (i) optimality of our proposed schemes for a certain class of straggler patterns and (ii) improved performance for the case of i.i.d. stragglers. These are further validated by experiments, where we implement our schemes to train neural networks.

Attack of the Tails: Yes, You Really Can Backdoor Federated Learning

- Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, Dimitris Papailiopoulos
- abstract@[open-review](#): Due to its decentralized nature, Federated Learning (FL) lends itself to adversarial attacks in the form of backdoors during training. The goal of a backdoor is to corrupt the performance of the trained model on specific sub-tasks (e.g., by classifying green cars as frogs). A range of FL backdoor attacks have been introduced in the literature, but also methods to defend against them, and it is currently an open question whether FL systems can be tailored to be robust against backdoors. In this work, we provide evidence to the contrary. We first establish that, in the general case, robustness to backdoors implies model robustness to adversarial examples, a major open problem in itself. Furthermore, detecting the presence of a backdoor in a FL model is unlikely assuming first-order oracles or polynomial time. We couple our theoretical results with a new family of backdoor attacks, which we refer to as edge-case backdoors. An edge-case backdoor forces a model to misclassify on seemingly easy inputs that are however unlikely to be part of the training, or test data, i.e., they live on the tail of the input distribution. We explain how these edge-case backdoors can lead to unsavory failures and may have serious repercussions on fairness. We further exhibit that, with careful tuning at the side of the adversary, one can insert them across a range of machine learning tasks (e.g., image classification, OCR, text prediction, sentiment analysis), and bypass state-of-the-art defense mechanisms.

Certifiably Adversarially Robust Detection of Out-of-Distribution Data

- Julian Bitterwolf, Alexander Meinke, Matthias Hein
- abstract@[open-review](#): Deep neural networks are known to be overconfident when applied to out-of-distribution (OOD) inputs which clearly do not belong to any class. This is a problem in safety-critical applications since a reliable assessment of the uncertainty of a classifier is a key property, allowing to trigger human intervention or to transfer into a safe state. In this paper, we are aiming for certifiable worst case guarantees for OOD detection by enforcing not only low confidence at the OOD point but also in an ℓ_1 -ball around it. For this purpose, we use interval bound propagation (IBP) to upper bound the maximal confidence in the ℓ_1 -ball and minimize this upper bound during training time. We show that non-trivial bounds on the confidence for OOD data generalizing beyond the OOD dataset seen at training time are possible. Moreover, in contrast to certified adversarial robustness which typically comes with significant loss in prediction performance, certified guarantees for worst case OOD detection are possible without much loss in accuracy.

Domain Generalization via Entropy Regularization

- Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, Dacheng Tao

- abstract@[open-review](#): Domain generalization aims to learn from multiple source domains a predictive model that can generalize to unseen target domains. One essential problem in domain generalization is to learn discriminative domain-invariant features. To arrive at this, some methods introduce a domain discriminator through adversarial learning to match the feature distributions in multiple source domains. However, adversarial training can only guarantee that the learned features have invariant marginal distributions, while the invariance of conditional distributions is more important for prediction in new domains. To ensure the conditional invariance of learned features, we propose an entropy regularization term that measures the dependency between the learned features and the class labels. Combined with the typical task-related loss, e.g., cross-entropy loss for classification, and adversarial loss for domain discrimination, our overall objective is guaranteed to learn conditional-invariant features across all source domains and thus can learn classifiers with better generalization capabilities. We demonstrate the effectiveness of our method through comparison with state-of-the-art methods on both simulated and real-world datasets. Code is available at: <https://github.com/sshan-zhao/DGviaER>.

Bayesian Meta-Learning for the Few-Shot Setting via Deep Kernels

- Massimiliano Patacchiola, Jack Turner, Elliot J. Crowley, Michael O'Boyle, Amos J. Storkey
- abstract@[open-review](#): Recently, different machine learning methods have been introduced to tackle the challenging few-shot learning scenario that is, learning from a small labeled dataset related to a specific task. Common approaches have taken the form of meta-learning: learning to learn on the new problem given the old. Following the recognition that meta-learning is implementing learning in a multi-level model, we present a Bayesian treatment for the meta-learning inner loop through the use of deep kernels. As a result we can learn a kernel that transfers to new tasks; we call this Deep Kernel Transfer (DKT). This approach has many advantages: is straightforward to implement as a single optimizer, provides uncertainty quantification, and does not require estimation of task-specific parameters. We empirically demonstrate that DKT outperforms several state-of-the-art algorithms in few-shot classification, and is the state of the art for cross-domain adaptation and regression. We conclude that complex meta-learning routines can be replaced by a simpler Bayesian model without loss of accuracy.

Skeleton-bridged Point Completion: From Global Inference to Local Adjustment

- Yinyu Nie, Yiqun Lin, Xiaoguang Han, Shihui Guo, Jian Chang, Shuguang Cui, Jian.J Zhang
- abstract@[open-review](#): Point completion refers to complete the missing geometries of objects from partial point clouds. Existing works usually estimate the missing shape by decoding a latent feature encoded from the input points. However, real-world objects are usually with diverse topologies and surface details, which a latent feature may fail to represent to recover a clean and complete surface. To this end, we propose a skeleton-bridged point completion network (SK-PCN) for shape completion. Given a partial scan, our method first predicts its 3D skeleton to obtain the global structure, and completes the surface by learning displacements from skeletal points. We decouple the shape completion into structure estimation and surface reconstruction, which eases the learning difficulty and benefits our method to obtain on-surface details. Besides, considering the missing features during encoding input points, SK-PCN adopts a local adjustment strategy that merges the input point cloud to our predictions for surface refinement. Comparing with previous methods, our skeleton-bridged manner better supports point normal estimation to obtain the full surface mesh beyond point clouds. The qualitative and quantitative experiments on both point cloud and mesh completion show that our approach outperforms the existing methods on various object categories.

Compressing Images by Encoding Their Latent Representations with Relative Entropy Coding

- Gergely Flamich, Marton Havasi, JosÃ© Miguel HernÃ¡ndez-Lobato
- abstract@[open-review](#): Variational Autoencoders (VAEs) have seen widespread use in learned image compression. They are used to learn expressive latent representations on which downstream compression methods can operate with high efficiency. Recently proposed 'bits-back' methods can indirectly encode the latent representation of images with codelength close to the relative entropy between the latent posterior and the prior. However, due to the underlying algorithm, these methods can only be used for lossless compression, and they only achieve their nominal efficiency when compressing multiple images simultaneously; they are inefficient for compressing single images. As an alternative, we propose a novel method, Relative Entropy Coding (REC), that can directly encode the latent representation with codelength close to the relative entropy for single images, supported by our empirical results obtained on the Cifar10, ImageNet32 and Kodak datasets. Moreover, unlike previous bits-back methods, REC is immediately applicable to lossy compression, where it is competitive with the state-of-the-art on the Kodak dataset.

Improved Guarantees for k-means++ and k-means++ Parallel

- Konstantin Makarychev, Aravind Reddy, Liren Shan
- abstract@[open-review](#): In this paper, we study k-means++ and k-meansll, the two most popular algorithms for the classic k-means clustering problem. We provide novel analyses and show improved approximation and bi-criteria approximation guarantees for k-means++ and k-meansll. Our results give a better theoretical justification for why these algorithms perform extremely well in practice.

Sparse Spectrum Warped Input Measures for Nonstationary Kernel Learning

- Anthony Tompkins, Rafael Oliveira, Fabio T. Ramos
- abstract@[open-review](#): We establish a general form of explicit, input-dependent, measure-valued warpings for learning nonstationary kernels. While stationary kernels are ubiquitous and simple to use, they struggle to adapt to functions that vary in smoothness with respect to the input. The proposed learning algorithm warps inputs as conditional Gaussian measures that control the smoothness of a standard stationary kernel. This construction allows us to capture non-stationary patterns in the data and provides intuitive inductive bias. The resulting method is based on sparse spectrum Gaussian processes, enabling closed-form solutions, and is extensible to a stacked construction to capture more complex patterns. The method is extensively validated alongside related algorithms on synthetic and real world datasets. We demonstrate a remarkable efficiency in the number of parameters of the warping functions in learning problems with both small and large data regimes.

An Efficient Adversarial Attack for Tree Ensembles

- Chong Zhang, Huan Zhang, Cho-Jui Hsieh
- abstract@[open-review](#): We study the problem of efficient adversarial attacks on tree based ensembles such as gradient boosting decision trees (GBDTs) and random forests (RFs). Since these models are non-continuous step functions and gradient does not exist, most existing efficient adversarial attacks are not applicable. Although decision-based black-box attacks can be applied, they cannot utilize the special structure of trees. In our work, we transform the attack problem into a discrete search problem specially designed for tree ensembles, where the goal is to find a valid ``leaf tuple'' that leads to misclassification while having the shortest distance to the original input. With this formulation, we show that a simple yet effective greedy algorithm can be applied to iteratively optimize the adversarial example by moving the leaf tuple to its neighborhood within hamming distance 1. Experimental results on several large GBDT and RF models with up to hundreds of trees demonstrate that our method can be thousands of times faster than the previous mixed-integer linear programming (MILP) based approach, while also providing smaller (better) adversarial examples than decision-based black-box attacks on general $\|\cdot\|_p$ ($p=1, 2, \infty$) norm perturbations.

Learning Continuous System Dynamics from Irregularly-Sampled Partial Observations

- Zijie Huang, Yizhou Sun, Wei Wang

- abstract@[open-review](#): Many real-world systems, such as moving planets, can be considered as multi-agent dynamic systems, where objects interact with each other and co-evolve along with the time. Such dynamics is usually difficult to capture, and understanding and predicting the dynamics based on observed trajectories of objects become a critical research problem in many domains. Most existing algorithms, however, assume the observations are regularly sampled and all the objects can be fully observed at each sampling time, which is impractical for many applications. In this paper, we propose to learn system dynamics from irregularly-sampled and partial observations with underlying graph structure for the first time. To tackle the above challenge, we present LG-ODE, a latent ordinary differential equation generative model for modeling multi-agent dynamic system with known graph structure. It can simultaneously learn the embedding of high dimensional trajectories and infer continuous latent system dynamics. Our model employs a novel encoder parameterized by a graph neural network that can infer initial states in an unsupervised way from irregularly-sampled partial observations of structural objects and utilizes neuralODE to infer arbitrarily complex continuous-time latent dynamics. Experiments on motion capture, spring system, and charged particle datasets demonstrate the effectiveness of our approach.

[Online Bayesian Persuasion](#)

- Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Nicola Gatti
- abstract@[open-review](#): In Bayesian persuasion, an informed sender has to design a signaling scheme that discloses the right amount of information so as to influence the behavior of a self-interested receiver. This kind of strategic interaction is ubiquitous in real economic scenarios. However, the original model by Kamenica and Gentzkow makes some stringent assumptions which limit its applicability in practice. One of the most limiting assumptions is arguably that, in order to compute an optimal signaling scheme, the sender is usually required to know the receiver's utility function. In this paper, we relax this assumption through an online learning framework in which the sender faces a receiver with unknown type. At each round, the receiver's type is chosen adversarially from a finite set of possible types. We are interested in no-regret algorithms prescribing a signaling scheme at each round of the repeated interaction with performances close to that of the best-in-hindsight signaling scheme. First, we prove a hardness result on the per-iteration running time required to achieve the no-regret property. Then, we provide algorithms for the full and partial information model which exhibit regret sublinear in the number of rounds and polynomial in the parameters of the game.

[Robust Pre-Training by Adversarial Contrastive Learning](#)

- Ziyu Jiang, Tianlong Chen, Ting Chen, Zhangyang Wang
- abstract@[open-review](#): Recent work has shown that, when integrated with adversarial training, self-supervised pre-training can lead to state-of-the-art robustness. In this work, we improve robustness-aware self-supervised pre-training by learning representations that are consistent under both data augmentations and adversarial perturbations. Our approach leverages a recent contrastive learning framework, which learns representations by maximizing feature consistency under differently augmented views. This fits particularly well with the goal of adversarial robustness, as one cause of adversarial fragility is the lack of feature invariance, i.e., small input perturbations can result in undesirable large changes in features or even predicted labels. We explore various options to formulate the contrastive task, and demonstrate that by injecting adversarial perturbations, contrastive pre-training can lead to models that are both label-efficient and robust. We empirically evaluate the proposed Adversarial Contrastive Learning (ACL) and show it can consistently outperform existing methods. For example on the CIFAR-10 dataset, ACL outperforms the previous state-of-the-art unsupervised robust pre-training approach by 2.99% on robust accuracy and 2.14% on standard accuracy. We further demonstrate that ACL pre-training can improve semi-supervised adversarial training, even when only a few labeled examples are available. Our codes and pre-trained models have been released at: <https://github.com/VITA-Group/Adversarial-Contrastive-Learning>.

[Random Walk Graph Neural Networks](#)

- Giannis Nikolentzos, Michalis Vazirgiannis
- abstract@[open-review](#): In recent years, graph neural networks (GNNs) have become the de facto tool for performing machine learning tasks on graphs. Most GNNs belong to the family of message passing neural networks (MPNNs). These models employ an iterative neighborhood aggregation scheme to update vertex representations. Then, to compute vector representations of graphs, they aggregate the representations of the vertices using some permutation invariant function. One would expect the hidden layers of a GNN to be composed of parameters that take the form of graphs. However, this is not the case for MPNNs since their update procedure is parameterized by fully-connected layers. In this paper, we propose a more intuitive and transparent architecture for graph-structured data, so-called Random Walk Graph Neural Network (RWNN). The first layer of the model consists of a number of trainable ``hidden graphs'' which are compared against the input graphs using a random walk kernel to produce graph representations. These representations are then passed on to a fully-connected neural network which produces the output. The employed random walk kernel is differentiable, and therefore, the proposed model is end-to-end trainable. We demonstrate the model's transparency on synthetic datasets. Furthermore, we empirically evaluate the model on graph classification datasets and show that it achieves competitive performance.

[Explore Aggressively, Update Conservatively: Stochastic Extragradient Methods with Variable Stepsize Scaling](#)

- Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, Panayotis Mertikopoulos
- abstract@[open-review](#): Owing to their stability and convergence speed, extragradient methods have become a staple for solving large-scale saddle-point problems in machine learning. The basic premise of these algorithms is the use of an extrapolation step before performing an update; thanks to this exploration step, extra-gradient methods overcome many of the non-convergence issues that plague gradient descent/ascent schemes. On the other hand, as we show in this paper, running vanilla extragradient with stochastic gradients may jeopardize its convergence, even in simple bilinear models. To overcome this failure, we investigate a double stepsize extragradient algorithm where the exploration step evolves at a more aggressive time-scale compared to the update step. We show that this modification allows the method to converge even with stochastic gradients, and we derive sharp convergence rates under an error bound condition.

[Fast and Accurate \\$k\\$-means++ via Rejection Sampling](#)

- Vincent Cohen-Addad, Silvio Lattanzi, Ashkan Norouzi-Fard, Christian Sohler, Ola Svensson
- abstract@[open-review](#): \$k\$-means++ \cite{arthur2007k} is a widely used clustering algorithm that is easy to implement, has nice theoretical guarantees and strong empirical performance. Despite its wide adoption, \$k\$-means++ sometimes suffers from being slow on large data-sets so a natural question has been to obtain more efficient algorithms with similar guarantees. In this paper, we present such a near linear time algorithm for \$k\$-means++ seeding. Interestingly our algorithm obtains the same theoretical guarantees as \$k\$-means++ and significantly improves earlier results on fast \$k\$-means++ seeding. Moreover, we show empirically that our algorithm is significantly faster than \$k\$-means++ and obtains solutions of equivalent quality.

[Variational Amodal Object Completion](#)

- Huan Ling, David Acuna, Karsten Kreis, Seung Wook Kim, Sanja Fidler
- abstract@[open-review](#): In images of complex scenes, objects are often occluding each other which makes perception tasks such as object detection and tracking, or robotic control tasks such as planning, challenging. To facilitate downstream tasks, it is thus important to reason about the full extent of objects, i.e., seeing behind occlusion, typically referred to as amodal instance completion. In this paper, we propose a variational generative framework for amodal completion, referred to as AMODAL-VAE, which does not require any amodal labels at training time, as it is able to utilize widely available object instance masks. We showcase our approach on the downstream task of scene editing where the user is presented with interactive tools to complete and erase objects in photographs. Experiments on complex street scenes demonstrate state-of-the-art performance in amodal mask completion and

showcase high-quality scene editing results. Interestingly, a user study shows that humans prefer object completions inferred by our model to the human-labeled ones.

[When Counterpoint Meets Chinese Folk Melodies](#)

- Nan Jiang, Sheng Jin, Zhiyao Duan, Changshui Zhang
- abstract@[open-review](#): Counterpoint is an important concept in Western music theory. In the past century, there have been significant interests in incorporating counterpoint into Chinese folk music composition. In this paper, we propose a reinforcement learning-based system, named FolkDuet, towards the online countermelody generation for Chinese folk melodies. With no existing data of Chinese folk duets, FolkDuet employs two reward models based on out-of-domain data, i.e. Bach chorales, and monophonic Chinese folk melodies. An interaction reward model is trained on the duets formed from outer parts of Bach chorales to model counterpoint interaction, while a style reward model is trained on monophonic melodies of Chinese folk songs to model melodic patterns. With both rewards, the generator of FolkDuet is trained to generate countermelodies while maintaining the Chinese folk style. The entire generation process is performed in an online fashion, allowing real-time interactive human-machine duet improvisation. Experiments show that the proposed algorithm achieves better subjective and objective results than the baselines.

[Sub-linear Regret Bounds for Bayesian Optimisation in Unknown Search Spaces](#)

- Hung Tran-The, Sunil Gupta, Santu Rana, Huong Ha, Svetha Venkatesh
- abstract@[open-review](#): Bayesian optimisation is a popular method for efficient optimisation of expensive black-box functions. Traditionally, BO assumes that the search space is known. However, in many problems, this assumption does not hold. To this end, we propose a novel BO algorithm which expands (and shifts) the search space over iterations based on controlling the expansion rate thought a \textbf{hyperharmonic series}. Further, we propose another variant of our algorithm that scales to high dimensions. We show theoretically that for both our algorithms, the cumulative regret grows at sub-linear rates. Our experiments with synthetic and real-world optimisation tasks demonstrate the superiority of our algorithms over the current state-of-the-art methods for Bayesian optimisation in unknown search space.

[Universal Domain Adaptation through Self Supervision](#)

- Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Kate Saenko
- abstract@[open-review](#): Unsupervised domain adaptation methods traditionally assume that all source categories are present in the target domain. In practice, little may be known about the category overlap between the two domains. While some methods address target settings with either partial or open-set categories, they assume that the particular setting is known a priori. We propose a more universally applicable domain adaptation approach that can handle arbitrary category shift, called Domain Adaptive Neighborhood Clustering via Entropy optimization (DANCE). Our approach combines two novel ideas: First, as we cannot fully rely on source categories to learn features discriminative for the target, we propose a novel neighborhood clustering technique to learn the structure of the target domain in a self-supervised way. Second, we use entropy-based feature alignment and rejection to align target features with the source, or reject them as unknown categories based on their entropy. We show through extensive experiments that DANCE outperforms baselines across open-set, open-partial and partial domain adaptation settings.

[Patch2Self: Denoising Diffusion MRI with Self-Supervised Learning](#)

- Shreyas Fadnavis, Joshua Batson, Eleftherios Garyfallidis
- abstract@[open-review](#): Diffusion-weighted magnetic resonance imaging (DWI) is the only non-invasive method for quantifying microstructure and reconstructing white-matter pathways in the living human brain. Fluctuations from multiple sources create significant noise in DWI data which must be suppressed before subsequent microstructure analysis. We introduce a self-supervised learning method for denoising DWI data, Patch2Self, which uses the entire volume to learn a full-rank locally linear denoiser for that volume. By taking advantage of the oversampled q-space of DWI data, Patch2Self can separate structure from noise without requiring an explicit model for either. We demonstrate the effectiveness of Patch2Self via quantitative and qualitative improvements in microstructure modeling, tracking (via fiber bundle coherency) and model estimation relative to other unsupervised methods on real and simulated data.

[Stochastic Normalization](#)

- Zhi Kou, Kaichao You, Mingsheng Long, Jianmin Wang
- abstract@[open-review](#): Fine-tuning pre-trained deep networks on a small dataset is an important component in the deep learning pipeline. A critical problem in fine-tuning is how to avoid over-fitting when data are limited. Existing efforts work from two aspects: (1) impose regularization on parameters or features; (2) transfer prior knowledge to fine-tuning by reusing pre-trained parameters. In this paper, we take an alternative approach by refactoring the widely used Batch Normalization (BN) module to mitigate over-fitting. We propose a two-branch design with one branch normalized by mini-batch statistics and the other branch normalized by moving statistics. During training, two branches are stochastically selected to avoid over-depending on some sample statistics, resulting in a strong regularization effect, which we interpret as ``architecture regularization.'' The resulting method is dubbed stochastic normalization (\textbf{StochNorm}). With the two-branch architecture, it naturally incorporates pre-trained moving statistics in BN layers during fine-tuning, exploiting more prior knowledge of pre-trained networks. Extensive empirical experiments show that StochNorm is a powerful tool to avoid over-fitting in fine-tuning with small datasets. Besides, StochNorm is readily pluggable in modern CNN backbones. It is complementary to other fine-tuning methods and can work together to achieve stronger regularization effect.

[Constrained episodic reinforcement learning in concave-convex and knapsack settings](#)

- Kiantâ© Brantley, Miro Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, Wen Sun
- abstract@[open-review](#): We propose an algorithm for tabular episodic reinforcement learning with constraints. We provide a modular analysis with strong theoretical guarantees for settings with concave rewards and convex constraints, and for settings with hard constraints (knapsacks). Most of the previous work in constrained reinforcement learning is limited to linear constraints, and the remaining work focuses on either the feasibility question or settings with a single episode. Our experiments demonstrate that the proposed algorithm significantly outperforms these approaches in existing constrained episodic environments.

[On Learning Ising Models under Huber's Contamination Model](#)

- Adarsh Prasad, Vishwakanth Srinivasan, Sivaraman Balakrishnan, Pradeep Ravikumar
- abstract@[open-review](#): We study the problem of learning Ising models in a setting where some of the samples from the underlying distribution can be arbitrarily corrupted. In such a setup, we aim to design statistically optimal estimators in a high-dimensional scaling in which the number of nodes p , the number of edges k and the maximal node degree d are allowed to increase to infinity as a function of the sample size n . Our analysis is based on exploiting moments of the underlying distribution, coupled with novel reductions to univariate estimation. Our proposed estimators achieve an optimal dimension independent dependence on the fraction of corrupted data in the contaminated setting, while also simultaneously achieving high-probability error guarantees with optimal sample-complexity. We corroborate our theoretical results by simulations.

[Cross-validation Confidence Intervals for Test Error](#)

- Pierre Bayle, Alexandre Bayle, Lucas Janson, Lester Mackey
- abstract@[open-review](#): This work develops central limit theorems for cross-validation and consistent estimators of the asymptotic variance under weak stability conditions on the learning algorithm. Together, these results provide practical, asymptotically-exact confidence intervals for k-fold test error and valid, powerful hypothesis tests of whether one learning algorithm has smaller k-fold test error than another. These results are also the first of their kind for the popular choice of leave-one-out cross-validation. In our experiments with diverse learning algorithms, the resulting intervals and tests outperform the most popular alternative methods from the literature.

[DeepSVG: A Hierarchical Generative Network for Vector Graphics Animation](#)

- Alexandre Carlier, Martin Danelljan, Alexandre Alahi, Radu Timofte
- abstract@[open-review](#): Scalable Vector Graphics (SVG) are ubiquitous in modern 2D interfaces due to their ability to scale to different resolutions. However, despite the success of deep learning-based models applied to rasterized images, the problem of vector graphics representation learning and generation remains largely unexplored. In this work, we propose a novel hierarchical generative network, called DeepSVG, for complex SVG icons generation and interpolation. Our architecture effectively disentangles high-level shapes from the low-level commands that encode the shape itself. The network directly predicts a set of shapes in a non-autoregressive fashion. We introduce the task of complex SVG icons generation by releasing a new large-scale dataset along with an open-source library for SVG manipulation. We demonstrate that our network learns to accurately reconstruct diverse vector graphics, and can serve as a powerful animation tool by performing interpolations and other latent space operations. Our code is available at <https://github.com/alexandre01/deepsvg>.

[Bayesian Attention Modules](#)

- Xinjie Fan, Shujian Zhang, Bo Chen, Mingyuan Zhou
- abstract@[open-review](#): Attention modules, as simple and effective tools, have not only enabled deep neural networks to achieve state-of-the-art results in many domains, but also enhanced their interpretability. Most current models use deterministic attention modules due to their simplicity and ease of optimization. Stochastic counterparts, on the other hand, are less popular despite their potential benefits. The main reason is that stochastic attention often introduces optimization issues or requires significant model changes. In this paper, we propose a scalable stochastic version of attention that is easy to implement and optimize. We construct simplex-constrained attention distributions by normalizing reparameterizable distributions, making the training process differentiable. We learn their parameters in a Bayesian framework where a data-dependent prior is introduced for regularization. We apply the proposed stochastic attention modules to various attention-based models, with applications to graph node classification, visual question answering, image captioning, machine translation, and language understanding. Our experiments show the proposed method brings consistent improvements over the corresponding baselines.

[Robustness Analysis of Non-Convex Stochastic Gradient Descent using Biased Expectations](#)

- Kevin Scaman, Cedric Malherbe
- abstract@[open-review](#): This work proposes a novel analysis of stochastic gradient descent (SGD) for non-convex and smooth optimization. Our analysis sheds light on the impact of the probability distribution of the gradient noise on the convergence rate of the norm of the gradient. In the case of sub-Gaussian and centered noise, we prove that, with probability $1-\delta$, the number of iterations to reach a precision ϵ for the squared gradient norm is $O(\epsilon^{-2}\ln(1/\delta))$. In the case of centered and integrable heavy-tailed noise, we show that, while the expectation of the iterates may be infinite, the squared gradient norm still converges with probability $1-\delta$ in $O(\epsilon^{-p}\delta^{-q})$ iterations, where $p,q > 2$. This result shows that heavy-tailed noise on the gradient slows down the convergence of SGD without preventing it, proving that SGD is robust to gradient noise with unbounded variance, a setting of interest for Deep Learning. In addition, it indicates that choosing a step size proportional to $T^{-1/b}$ where b is the tail-parameter of the noise and T is the number of iterations leads to the best convergence rates. Both results are simple corollaries of a unified analysis using the novel concept of biased expectations, a simple and intuitive mathematical tool to obtain concentration inequalities. Using this concept, we propose a new quantity to measure the amount of noise added to the gradient, and discuss its value in multiple scenarios.

[SoftFlow: Probabilistic Framework for Normalizing Flow on Manifolds](#)

- Hyeongju Kim, Hyeonseung Lee, Woo Hyun Kang, Joun Yeop Lee, Nam Soo Kim
- abstract@[open-review](#): Flow-based generative models are composed of invertible transformations between two random variables of the same dimension. Therefore, flow-based models cannot be adequately trained if the dimension of the data distribution does not match that of the underlying target distribution. In this paper, we propose SoftFlow, a probabilistic framework for training normalizing flows on manifolds. To sidestep the dimension mismatch problem, SoftFlow estimates a conditional distribution of the perturbed input data instead of learning the data distribution directly. We experimentally show that SoftFlow can capture the innate structure of the manifold data and generate high-quality samples unlike the conventional flow-based models. Furthermore, we apply the proposed framework to 3D point clouds to alleviate the difficulty of forming thin structures for flow-based models. The proposed model for 3D point clouds, namely SoftPointFlow, can estimate the distribution of various shapes more accurately and achieves state-of-the-art performance in point cloud generation.

[A meta-learning approach to \(re\)discover plasticity rules that carve a desired function into a neural network](#)

- Basile Confavreux, Friedemann Zenke, Everton Agnes, Timothy Lillicrap, Tim Vogels
- abstract@[open-review](#): The search for biologically faithful synaptic plasticity rules has resulted in a large body of models. They are usually inspired by -- and fitted to -- experimental data, but they rarely produce neural dynamics that serve complex functions. These failures suggest that current plasticity models are still under-constrained by existing data. Here, we present an alternative approach that uses meta-learning to discover plausible synaptic plasticity rules. Instead of experimental data, the rules are constrained by the functions they implement and the structure they are meant to produce. Briefly, we parameterize synaptic plasticity rules by a Volterra expansion and then use supervised learning methods (gradient descent or evolutionary strategies) to minimize a problem-dependent loss function that quantifies how effectively a candidate plasticity rule transforms an initially random network into one with the desired function. We first validate our approach by re-discovering previously described plasticity rules, starting at the single-neuron level and ``Oja's rule'', a simple Hebbian plasticity rule that captures the direction of most variability of inputs to a neuron (i.e., the first principal component). We expand the problem to the network level and ask the framework to find Oja's rule together with an anti-Hebbian rule such that an initially random two-layer firing-rate network will recover several principal components of the input space after learning. Next, we move to networks of integrate-and-fire neurons with plastic inhibitory afferents. We train for rules that achieve a target firing rate by countering tuned excitation. Our algorithm discovers a specific subset of the manifold of rules that can solve this task. Our work is a proof of principle of an automated and unbiased approach to unveil synaptic plasticity rules that obey biological constraints and can solve complex functions.

[Greedy Optimization Provably Wins the Lottery: Logarithmic Number of Winning Tickets is Enough](#)

- Mao Ye, Lemeng Wu, Qiang Liu

- abstract@[open-review](#): Despite the great success of deep learning, recent works show that large deep neural networks are often highly redundant and can be significantly reduced in size. However, the theoretical question of how much we can prune a neural network given a specified tolerance of accuracy drop is still open. This paper provides one answer to this question by proposing a greedy optimization based pruning method. The proposed method has the guarantee that the discrepancy between the pruned network and the original network decays with exponentially fast rate w.r.t. the size of the pruned network, under weak assumptions that apply for most practical settings. Empirically, our method improves prior arts on pruning various network architectures including ResNet, MobilenetV2/V3 on ImageNet.

[Path Integral Based Convolution and Pooling for Graph Neural Networks](#)

- Zheng Ma, Junyu Xuan, Yu Guang Wang, Ming Li, Pietro Li \tilde{A} ²
- abstract@[open-review](#): Graph neural networks (GNNs) extends the functionality of traditional neural networks to graph-structured data. Similar to CNNs, an optimized design of graph convolution and pooling is key to success. Borrowing ideas from physics, we propose a path integral based graph neural networks (PAN) for classification and regression tasks on graphs. Specifically, we consider a convolution operation that involves every path linking the message sender and receiver with learnable weights depending on the path length, which corresponds to the maximal entropy random walk. It generalizes the graph Laplacian to a new transition matrix we call \emph{maximal entropy transition} (MET) matrix derived from a path integral formalism. Importantly, the diagonal entries of the MET matrix are directly related to the subgraph centrality, thus lead to a natural and adaptive pooling mechanism. PAN provides a versatile framework that can be tailored for different graph data with varying sizes and structures. We can view most existing GNN architectures as special cases of PAN. Experimental results show that PAN achieves state-of-the-art performance on various graph classification/regression tasks, including a new benchmark dataset from statistical mechanics we propose to boost applications of GNN in physical sciences.

[Estimating the Effects of Continuous-valued Interventions using Generative Adversarial Networks](#)

- Ioana Bica, James Jordon, Mihaela van der Schaar
- abstract@[open-review](#): While much attention has been given to the problem of estimating the effect of discrete interventions from observational data, relatively little work has been done in the setting of continuous-valued interventions, such as treatments associated with a dosage parameter. In this paper, we tackle this problem by building on a modification of the generative adversarial networks (GANs) framework. Our model, SCIGAN, is flexible and capable of simultaneously estimating counterfactual outcomes for several different continuous interventions. The key idea is to use a significantly modified GAN model to learn to generate counterfactual outcomes, which can then be used to learn an inference model, using standard supervised methods, capable of estimating these counterfactuals for a new sample. To address the challenges presented by shifting to continuous interventions, we propose a novel architecture for our discriminator - we build a hierarchical discriminator that leverages the structure of the continuous intervention setting. Moreover, we provide theoretical results to support our use of the GAN framework and of the hierarchical discriminator. In the experiments section, we introduce a new semi-synthetic data simulation for use in the continuous intervention setting and demonstrate improvements over the existing benchmark models.

[Latent Dynamic Factor Analysis of High-Dimensional Neural Recordings](#)

- Heejong Bong, Zongge Liu, Zhao Ren, Matthew Smith, Valerie Ventura, Robert E. Kass
- abstract@[open-review](#): High-dimensional neural recordings across multiple brain regions can be used to establish functional connectivity with good spatial and temporal resolution. We designed and implemented a novel method, Latent Dynamic Factor Analysis of High-dimensional time series (LDFA-H), which combines (a) a new approach to estimating the covariance structure among high-dimensional time series (for the observed variables) and (b) a new extension of probabilistic CCA to dynamic time series (for the latent variables). Our interest is in the cross-correlations among the latent variables which, in neural recordings, may capture the flow of information from one brain region to another. Simulations show that LDFA-H outperforms existing methods in the sense that it captures target factors even when within-region correlation due to noise dominates cross-region correlation. We applied our method to local field potential (LFP) recordings from 192 electrodes in Prefrontal Cortex (PFC) and visual area V4 during a memory-guided saccade task. The results capture time-varying lead-lag dependencies between PFC and V4, and display the associated spatial distribution of the signals.

[Conditioning and Processing: Techniques to Improve Information-Theoretic Generalization Bounds](#)

- Hassan Hafez-Kolahi, Zeinab Golgooni, Shohreh Kasaei, Mahdieh Soleymani
- abstract@[open-review](#): Obtaining generalization bounds for learning algorithms is one of the main subjects studied in theoretical machine learning. In recent years, information-theoretic bounds on generalization have gained the attention of researchers. This approach provides an insight into learning algorithms by considering the mutual information between the model and the training set. In this paper, a probabilistic graphical representation of this approach is adopted and two general techniques to improve the bounds are introduced, namely conditioning and processing. In conditioning, a random variable in the graph is considered as given, while in processing a random variable is substituted with one of its children. These techniques can be used to improve the bounds by either sharpening them or increasing their applicability. It is demonstrated that the proposed framework provides a simple and unified way to explain a variety of recent tightening results. New improved bounds derived by utilizing these techniques are also proposed.

[Bongard-LOGO: A New Benchmark for Human-Level Concept Learning and Reasoning](#)

- Weili Nie, Zhiding Yu, Lei Mao, Ankit B. Patel, Yuke Zhu, Anima Anandkumar
- abstract@[open-review](#): Humans have an inherent ability to learn novel concepts from only a few samples and generalize these concepts to different situations. Even though today's machine learning models excel with a plethora of training data on standard recognition tasks, a considerable gap exists between machine-level pattern recognition and human-level concept learning. To narrow this gap, the Bongard Problems (BPs) were introduced as an inspirational challenge for visual cognition in intelligent systems. Albeit new advances in representation learning and learning to learn, BPs remain a daunting challenge for modern AI. Inspired by the original one hundred BPs, we propose a new benchmark Bongard-LOGO for human-level concept learning and reasoning. We develop a program-guided generation technique to produce a large set of human-interpretable visual cognition problems in action-oriented LOGO language. Our benchmark captures three core properties of human cognition: 1) context-dependent perception, in which the same object may have disparate interpretations given different contexts; 2) analogy-making perception, in which some meaningful concepts are traded off for other meaningful concepts; and 3) perception with a few samples but infinite vocabulary. In experiments, we show that the state-of-the-art deep learning methods perform substantially worse than human subjects, implying that they fail to capture core human cognition properties. Finally, we discuss research directions towards a general architecture for visual reasoning to tackle this benchmark.

[GAN Memory with No Forgetting](#)

- Yulai Cong, Miaoyun Zhao, Jianqiao Li, Sijia Wang, Lawrence Carin
- abstract@[open-review](#): As a fundamental issue in lifelong learning, catastrophic forgetting is directly caused by inaccessible historical data; accordingly, if the data (information) were memorized perfectly, no forgetting should be expected. Motivated by that, we propose a GAN memory for lifelong learning, which is capable of remembering a stream of datasets via generative processes, with \emph{no} forgetting. Our GAN memory is based on recognizing that one can modulate the ``style'' of a GAN model to form perceptually-distant targeted generation. Accordingly, we propose to do sequential style modulations atop a well-behaved base GAN model, to form sequential targeted generative models, while simultaneously benefiting from the transferred base knowledge. The GAN memory -- that is motivated by lifelong learning -- is therefore itself manifested by a form of lifelong learning, via forward transfer and modulation of information from prior tasks. Experiments demonstrate the superiority of our method over existing approaches and its

effectiveness in alleviating catastrophic forgetting for lifelong classification problems. Code is available at \url{https://github.com/MiaoyunZhao/GANmemory_LifelongLearning}.

[Deep Reinforcement Learning with Stacked Hierarchical Attention for Text-based Games](#)

- Yunqiu Xu, Meng Fang, Ling Chen, Yali Du, Joey Tianyi Zhou, Chengqi Zhang
- abstract@[open-review](#): We study reinforcement learning (RL) for text-based games, which are interactive simulations in the context of natural language. While different methods have been developed to represent the environment information and language actions, existing RL agents are not empowered with any reasoning capabilities to deal with textual games. In this work, we aim to conduct explicit reasoning with knowledge graphs for decision making, so that the actions of an agent are generated and supported by an interpretable inference procedure. We propose a stacked hierarchical attention mechanism to construct an explicit representation of the reasoning process by exploiting the structure of the knowledge graph. We extensively evaluate our method on a number of man-made benchmark games, and the experimental results demonstrate that our method performs better than existing text-based agents.

[Gaussian Gated Linear Networks](#)

- David Budden, Adam Marblestone, Eren Sezener, Tor Lattimore, Gregory Wayne, Joel Veness
- abstract@[open-review](#): We propose the Gaussian Gated Linear Network (G-GLN), an extension to the recently proposed GLN family of deep neural networks. Instead of using backpropagation to learn features, GLNs have a distributed and local credit assignment mechanism based on optimizing a convex objective. This gives rise to many desirable properties including universality, data-efficient online learning, trivial interpretability and robustness to catastrophic forgetting. We extend the GLN framework from classification to multiple regression and density modelling by generalizing geometric mixing to a product of Gaussian densities. The G-GLN achieves competitive or state-of-the-art performance on several univariate and multivariate regression benchmarks, and we demonstrate its applicability to practical tasks including online contextual bandits and density estimation via denoising.

[Node Classification on Graphs with Few-Shot Novel Labels via Meta Transformed Network Embedding](#)

- Lin Lan, Pingui Wang, Xuefeng Du, Kaikai Song, Jing Tao, Xiaohong Guan
- abstract@[open-review](#): We study the problem of node classification on graphs with few-shot novel labels, which has two distinctive properties: (1) There are novel labels to emerge in the graph; (2) The novel labels have only a few representative nodes for training a classifier. The study of this problem is instructive and corresponds to many applications such as recommendations for newly formed groups with only a few users in online social networks. To cope with this problem, we propose a novel Meta Transformed Network Embedding framework (MetaTNE), which consists of three modules: (1) A \emph{structural module} provides each node a latent representation according to the graph structure. (2) A \emph{meta-learning module} captures the relationships between the graph structure and the node labels as prior knowledge in a meta-learning manner. Additionally, we introduce an \emph{embedding transformation function} that remedies the deficiency of the straightforward use of meta-learning. Inherently, the meta-learned prior knowledge can be used to facilitate the learning of few-shot novel labels. (3) An \emph{optimization module} employs a simple yet effective scheduling strategy to train the above two modules with a balance between graph structure learning and meta-learning. Experiments on four real-world datasets show that MetaTNE brings a huge improvement over the state-of-the-art methods.

[Online Fast Adaptation and Knowledge Accumulation \(OSAKA\): a New Approach to Continual Learning](#)

- Massimo Caccia, Pau Rodriguez, Oleksiy Ostapenko, Fabrice Normandin, Min Lin, Lucas Page-Caccia, Issam Hadj Laradji, Irina Rish, Alexandre Lacoste, David Vázquez, Laurent Charlin
- abstract@[open-review](#): Continual learning agents experience a stream of (related) tasks. The main challenge is that the agent must not forget previous tasks and also adapt to novel tasks in the stream. We are interested in the intersection of two recent continual-learning scenarios. In meta-continual learning, the model is pre-trained using meta-learning to minimize catastrophic forgetting of previous tasks. In continual-meta learning, the aim is to train agents for faster remembering of previous tasks through adaptation. In their original formulations, both methods have limitations. We stand on their shoulders to propose a more general scenario, OSAKA, where an agent must quickly solve new (out-of-distribution) tasks, while also requiring fast remembering. We show that current continual learning, meta-learning, meta-continual learning, and continual-meta learning techniques fail in this new scenario. We propose Continual-MAML, an online extension of the popular MAML algorithm as a strong baseline for this scenario. We show in an empirical study that Continual-MAML is better suited to the new scenario than the aforementioned methodologies including standard continual learning and meta-learning approaches.

[Convex optimization based on global lower second-order models](#)

- Nikita Doikov, Yurii Nesterov
- abstract@[open-review](#): In this work, we present new second-order algorithms for composite convex optimization, called Contracting-domain Newton methods. These algorithms are affine-invariant and based on global second-order lower approximation for the smooth component of the objective. Our approach has an interpretation both as a second-order generalization of the conditional gradient method, or as a variant of trust-region scheme. Under the assumption, that the problem domain is bounded, we prove $\mathcal{O}(1/k^2)$ global rate of convergence in functional residual, where k is the iteration counter, minimizing convex functions with Lipschitz continuous Hessian. This significantly improves the previously known bound $\mathcal{O}(1/k)$ for this type of algorithms. Additionally, we propose a stochastic extension of our method, and present computational results for solving empirical risk minimization problem.

[Simultaneously Learning Stochastic and Adversarial Episodic MDPs with Known Transition](#)

- Tiancheng Jin, Haipeng Luo
- abstract@[open-review](#): This work studies the problem of learning episodic Markov Decision Processes with known transition and bandit feedback. We develop the first algorithm with a ``best-of-both-worlds'' guarantee: it achieves $O(\log T)$ regret when the losses are stochastic, and simultaneously enjoys worst-case robustness with $\tilde{O}(\sqrt{T})$ regret even when the losses are adversarial, where T is the number of episodes. More generally, it achieves $\tilde{O}(\sqrt{C})$ regret in an intermediate setting where the losses are corrupted by a total amount of C . Our algorithm is based on the Follow-the-Regularized-Leader method from Zimin and Neu (2013), with a novel hybrid regularizer inspired by recent works of Zimmert et al. (2019a, 2019b) for the special case of multi-armed bandits. Crucially, our regularizer admits a non-diagonal Hessian with a highly complicated inverse. Analyzing such a regularizer and deriving a particular self-bounding regret guarantee is our key technical contribution and might be of independent interest.

[Relative gradient optimization of the Jacobian term in unsupervised deep learning](#)

- Luigi Gresele, Giancarlo Fissore, Adrián Javaloy, Bernhard Schölkopf, Aapo Hyvärinen
- abstract@[open-review](#): Learning expressive probabilistic models correctly describing the data is a ubiquitous problem in machine learning. A popular approach for solving it is mapping the observations into a representation space with a simple joint distribution, which can typically be written as a product of its marginals â€” thus drawing a connection with the field of nonlinear independent component analysis. Deep density models have been widely used for this task, but their maximum likelihood based training requires estimating the log-determinant of the Jacobian and is computationally expensive, thus imposing a trade-off between computation and expressive power. In this work, we propose a new approach for exact training of such neural networks.

Based on relative gradients, we exploit the matrix structure of neural network parameters to compute updates efficiently even in high-dimensional spaces; the computational cost of the training is quadratic in the input size, in contrast with the cubic scaling of naive approaches. This allows fast training with objective functions involving the log-determinant of the Jacobian, without imposing constraints on its structure, in stark contrast to autoregressive normalizing flows.

[Self-Supervised Visual Representation Learning from Hierarchical Grouping](#)

- Xiao Zhang, Michael Maire
- abstract@[open-review](#): We create a framework for bootstrapping visual representation learning from a primitive visual grouping capability. We operationalize grouping via a contour detector that partitions an image into regions, followed by merging of those regions into a tree hierarchy. A small supervised dataset suffices for training this grouping primitive. Across a large unlabeled dataset, we apply this learned primitive to automatically predict hierarchical region structure. These predictions serve as guidance for self-supervised contrastive feature learning: we task a deep network with producing per-pixel embeddings whose pairwise distances respect the region hierarchy. Experiments demonstrate that our approach can serve as state-of-the-art generic pre-training, benefiting downstream tasks. We additionally explore applications to semantic region search and video-based object instance tracking.

[Optimal Variance Control of the Score-Function Gradient Estimator for Importance-Weighted Bounds](#)

- Valentin Lićvin, Andrea Dittadi, Anders Christensen, Ole Winther
- abstract@[open-review](#): This paper introduces novel results for the score-function gradient estimator of the importance-weighted variational bound (IWAE). We prove that in the limit of large K (number of importance samples) one can choose the control variate such that the Signal-to-Noise ratio (SNR) of the estimator grows as \sqrt{K} . This is in contrast to the standard pathwise gradient estimator where the SNR decreases as $1/\sqrt{K}$. Based on our theoretical findings we develop a novel control variate that extends on VIMCO. Empirically, for the training of both continuous and discrete generative models, the proposed method yields superior variance reduction, resulting in an SNR for IWAE that increases with K without relying on the reparameterization trick. The novel estimator is competitive with state-of-the-art reparameterization-free gradient estimators such as Reweighted Wake-Sleep (RWS) and the thermodynamic variational objective (TVO) when training generative models.

[Explicit Regularisation in Gaussian Noise Injections](#)

- Alexander Camuto, Matthew Willetts, Umut Simsekli, Stephen J. Roberts, Chris C. Holmes
- abstract@[open-review](#): We study the regularisation induced in neural networks by Gaussian noise injections (GNIs). Though such injections have been extensively studied when applied to data, there have been few studies on understanding the regularising effect they induce when applied to network activations. Here we derive the explicit regulariser of GNIs, obtained by marginalising out the injected noise, and show that it penalises functions with high-frequency components in the Fourier domain; particularly in layers closer to a neural network's output. We show analytically and empirically that such regularisation produces calibrated classifiers with large classification margins.

[Numerically Solving Parametric Families of High-Dimensional Kolmogorov Partial Differential Equations via Deep Learning](#)

- Julius Berner, Markus Dablander, Philipp Grohs
- abstract@[open-review](#): We present a deep learning algorithm for the numerical solution of parametric families of high-dimensional linear Kolmogorov partial differential equations (PDEs). Our method is based on reformulating the numerical approximation of a whole family of Kolmogorov PDEs as a single statistical learning problem using the Feynman-Kac formula. Successful numerical experiments are presented, which empirically confirm the functionality and efficiency of our proposed algorithm in the case of heat equations and Black-Scholes option pricing models parametrized by affine-linear coefficient functions. We show that a single deep neural network trained on simulated data is capable of learning the solution functions of an entire family of PDEs on a full space-time region. Most notably, our numerical observations and theoretical results also demonstrate that the proposed method does not suffer from the curse of dimensionality, distinguishing it from almost all standard numerical methods for PDEs.

[Finite-Time Analysis for Double Q-learning](#)

- Huaqing Xiong, Lin Zhao, Yingbin Liang, Wei Zhang
- abstract@[open-review](#): Although Q-learning is one of the most successful algorithms for finding the best action-value function (and thus the optimal policy) in reinforcement learning, its implementation often suffers from large overestimation of Q-function values incurred by random sampling. The double Q-learning algorithm proposed in [hasselt2010double](#) overcomes such an overestimation issue by randomly switching the update between two Q-estimators, and has thus gained significant popularity in practice. However, the theoretical understanding of double Q-learning is rather limited. So far only the asymptotic convergence has been established, which does not characterize how fast the algorithm converges. In this paper, we provide the first non-asymptotic (i.e., finite-time) analysis for double Q-learning. We show that both synchronous and asynchronous double Q-learning are guaranteed to converge to an ϵ -accurate neighborhood of the global optimum by taking $\tilde{\Omega}(\frac{1}{(1-\gamma)^6\epsilon^2})^{1/\Omega}$ iterations, where $\Omega \in (0,1)$ is the decay parameter of the learning rate, and γ is the discount factor. Our analysis develops novel techniques to derive finite-time bounds on the difference between two inter-connected stochastic processes, which is new to the literature of stochastic approximation.

[Learning to Detect Objects with a 1 Megapixel Event Camera](#)

- Etienne Perot, Pierre de Tournemire, Davide Nitti, Jonathan Masci, Amos Sironi
- abstract@[open-review](#): Event cameras encode visual information with high temporal precision, low data-rate, and high-dynamic range. Thanks to these characteristics, event cameras are particularly suited for scenarios with high motion, challenging lighting conditions and requiring low latency. However, due to the novelty of the field, the performance of event-based systems on many vision tasks is still lower compared to conventional frame-based solutions. The main reasons for this performance gap are: the lower spatial resolution of event sensors, compared to frame cameras; the lack of large-scale training datasets; the absence of well established deep learning architectures for event-based processing. In this paper, we address all these problems in the context of an event-based object detection task. First, we publicly release the first high-resolution large-scale dataset for object detection. The dataset contains more than 14 hours recordings of a 1 megapixel event camera, in automotive scenarios, together with 25M bounding boxes of cars, pedestrians, and two-wheelers, labeled at high frequency. Second, we introduce a novel recurrent architecture for event-based detection and a temporal consistency loss for better-behaved training. The ability to compactly represent the sequence of events into the internal memory of the model is essential to achieve high accuracy. Our model outperforms by a large margin feed-forward event-based architectures. Moreover, our method does not require any reconstruction of intensity images from events, showing that training directly from raw events is possible, more efficient, and more accurate than passing through an intermediate intensity image. Experiments on the dataset introduced in this work, for which events and gray level images are available, show performance on par with that of highly tuned and studied frame-based detectors.

[End-to-End Learning and Intervention in Games](#)

- Jiayang Li, Jing Yu, Yu Nie, Zhaoran Wang

- abstract@[open-review](#): In a social system, the self-interest of agents can be detrimental to the collective good, sometimes leading to social dilemmas. To resolve such a conflict, a central designer may intervene by either redesigning the system or incentivizing the agents to change their behaviors. To be effective, the designer must anticipate how the agents react to the intervention, which is dictated by their often unknown payoff functions. Therefore, learning about the agents is a prerequisite for intervention. In this paper, we provide a unified framework for learning and intervention in games. We cast the equilibria of games as individual layers and integrate them into an end-to-end optimization framework. To enable the backward propagation through the equilibria of games, we propose two approaches, respectively based on explicit and implicit differentiation. Specifically, we cast the equilibria as the solutions to variational inequalities (VIs). The explicit approach unrolls the projection method for solving VIs, while the implicit approach exploits the sensitivity of the solutions to VIs. At the core of both approaches is the differentiation through a projection operator. Moreover, we establish the correctness of both approaches and identify the conditions under which one approach is more desirable than the other. The analytical results are validated using several real-world problems.

[Least Squares Regression with Markovian Data: Fundamental Limits and Algorithms](#)

- Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, Praneeth Netrapalli
- abstract@[open-review](#): We study the problem of least squares linear regression where the datapoints are dependent and are sampled from a Markov chain. We establish sharp information theoretic minimax lower bounds for this problem in terms of $\$tmix$$, the mixing time of the underlying Markov chain, under different noise settings. Our results establish that in general, optimization with Markovian data is strictly harder than optimization with independent data and a trivial algorithm (SGD-DD) that works with only one in every $\$tmix$$ samples, which are approximately independent, is minimax optimal. In fact, it is strictly better than the popular Stochastic Gradient Descent (SGD) method with constant step-size which is otherwise minimax optimal in the regression with independent data setting.

Beyond a worst case analysis, we investigate whether structured datasets seen in practice such as Gaussian auto-regressive dynamics can admit more efficient optimization schemes. Surprisingly, even in this specific and natural setting, Stochastic Gradient Descent (SGD) with constant step-size is still no better than SGD-DD. Instead, we propose an algorithm based on experience replay--a popular reinforcement learning technique--that achieves a significantly better error rate. Our improved rate serves as one of the first results where an algorithm outperforms SGD-DD on an interesting Markov chain and also provides one of the first theoretical analyses to support the use of experience replay in practice.

[Predictive coding in balanced neural networks with noise, chaos and delays](#)

- Jonathan Kadmon, Jonathan Timcheck, Surya Ganguli
- abstract@[open-review](#): Biological neural networks face a formidable task: performing reliable computations in the face of intrinsic stochasticity in individual neurons, imprecisely specified synaptic connectivity, and nonnegligible delays in synaptic transmission. A common approach to combatting such biological heterogeneity involves averaging over large redundant networks of N neurons resulting in coding errors that decrease classically as the square root of N . Recent work demonstrated a novel mechanism whereby recurrent spiking networks could efficiently encode dynamic stimuli achieving a superclassical scaling in which coding errors decrease as $1/N$. This specific mechanism involved two key ideas: predictive coding, and a tight balance, or cancellation between strong feedforward inputs and strong recurrent feedback. However, the theoretical principles governing the efficacy of balanced predictive coding and its robustness to noise, synaptic weight heterogeneity and communication delays remain poorly understood. To discover such principles, we introduce an analytically tractable model of balanced predictive coding, in which the degree of balance and the degree of weight disorder can be dissociated unlike in previous balanced network models, and we develop a mean-field theory of coding accuracy. Overall, our work provides and solves a general theoretical framework for dissecting the differential contributions neural noise, synaptic disorder, chaos, synaptic delays, and balance to the fidelity of predictive neural codes, reveals the fundamental role that balance plays in achieving superclassical scaling, and unifies previously disparate models in theoretical neuroscience.

[Interpolation Technique to Speed Up Gradients Propagation in Neural ODEs](#)

- Talgat Daulbaev, Alexandr Katrutsa, Larisa Markeeva, Julia Gusak, Andrzej Cichocki, Ivan Oseledets
- abstract@[open-review](#): We propose a simple interpolation-based method for the efficient approximation of gradients in neural ODE models. We compare it with reverse dynamic method (known in literature as “adjoint method”) to train neural ODEs on classification, density estimation and inference approximation tasks. We also propose a theoretical justification of our approach using logarithmic norm formalism. As a result, our method allows faster model training than the reverse dynamic method what was confirmed and validated by extensive numerical experiments for several standard benchmarks.

[On the Equivalence between Online and Private Learnability beyond Binary Classification](#)

- Young Jung, Baekjin Kim, Ambuj Tewari
- abstract@[open-review](#): Alon et al. [2019] and Bun et al. [2020] recently showed that online learnability and private PAC learnability are equivalent in binary classification. We investigate whether this equivalence extends to multi-class classification and regression. First, we show that private learnability implies online learnability in both settings. Our extension involves studying a novel variant of the Littlestone dimension that depends on a tolerance parameter and on an appropriate generalization of the concept of threshold functions beyond binary classification. Second, we show that while online learnability continues to imply private learnability in multi-class classification, current proof techniques encounter significant hurdles in the regression setting. While the equivalence for regression remains open, we provide non-trivial sufficient conditions for an online learnable class to also be privately learnable.

[AViD Dataset: Anonymized Videos from Diverse Countries](#)

- AJ Piergiovanni, Michael Ryoo
- abstract@[open-review](#): We introduce a new public video dataset for action recognition: Anonymized Videos from Diverse countries (AViD). Unlike existing public video datasets, AViD is a collection of action videos from many different countries. The motivation is to create a public dataset that would benefit training and pretraining of action recognition models for everybody, rather than making it useful for limited countries. Further, all the face identities in the AViD videos are properly anonymized to protect their privacy. It also is a static dataset where each video is licensed with the creative commons license. We confirm that most of the existing video datasets are statistically biased to only capture action videos from a limited number of countries. We experimentally illustrate that models trained with such biased datasets do not transfer perfectly to action videos from the other countries, and show that AViD addresses such problem. We also confirm that the new AViD dataset could serve as a good dataset for pretraining the models, performing comparably or better than prior datasets. The dataset is available at <https://github.com/piergaj/AViD>

[Probably Approximately Correct Constrained Learning](#)

- Luiz Chamon, Alejandro Ribeiro
- abstract@[open-review](#): As learning solutions reach critical applications in social, industrial, and medical domains, the need to curtail their behavior has become paramount. There is now ample evidence that without explicit tailoring, learning can lead to biased, unsafe, and prejudiced solutions. To tackle these problems, we develop a generalization theory of constrained learning based on the probably approximately correct (PAC) learning framework. In particular, we show that imposing requirements does not make a learning problem harder in the sense that any PAC learnable class is also PAC constrained learnable using a constrained counterpart of the empirical risk minimization (ERM) rule. For typical parametrized models, however, this learner involves

solving a constrained non-convex optimization program for which even obtaining a feasible solution is challenging. To overcome this issue, we prove that under mild conditions the empirical dual problem of constrained learning is also a PAC constrained learner that now leads to a practical constrained learning algorithm based solely on solving unconstrained problems. We analyze the generalization properties of this solution and use it to illustrate how constrained learning can address problems in fair and robust classification.

[RATT: Recurrent Attention to Transient Tasks for Continual Image Captioning](#)

- Riccardo Del Chiaro, Bartłomiej Twardowski, Andrew Bagdanov, Joost van de Weijer
- abstract@[open-review](#): Research on continual learning has led to a variety of approaches to mitigating catastrophic forgetting in feed-forward classification networks. Until now surprisingly little attention has been focused on continual learning of recurrent models applied to problems like image captioning. In this paper we take a systematic look at continual learning of LSTM-based models for image captioning. We propose an attention-based approach that explicitly accommodates the transient nature of vocabularies in continual image captioning tasks -- i.e. that task vocabularies are not disjoint. We call our method Recurrent Attention to Transient Tasks (RATT), and also show how to adapt continual learning approaches based on weight regularization and knowledge distillation to recurrent continual learning problems. We apply our approaches to incremental image captioning problem on two new continual learning benchmarks we define using the MS-COCO and Flickr30 datasets. Our results demonstrate that RATT is able to sequentially learn five captioning tasks while incurring no forgetting of previously learned ones.

[Decisions, Counterfactual Explanations and Strategic Behavior](#)

- Stratis Tsirtsis, Manuel Gomez Rodriguez
- abstract@[open-review](#): As data-driven predictive models are increasingly used to inform decisions, it has been argued that decision makers should provide explanations that help individuals understand what would have to change for these decisions to be beneficial ones. However, there has been little discussion on the possibility that individuals may use the above counterfactual explanations to invest effort strategically and maximize their chances of receiving a beneficial decision. In this paper, our goal is to find policies and counterfactual explanations that are optimal in terms of utility in such a strategic setting. We first show that, given a pre-defined policy, the problem of finding the optimal set of counterfactual explanations is NP-hard. Then, we show that the corresponding objective is nondecreasing and satisfies submodularity and this allows a standard greedy algorithm to enjoy approximation guarantees. In addition, we further show that the problem of jointly finding both the optimal policy and set of counterfactual explanations reduces to maximizing a non-monotone submodular function. As a result, we can use a recent randomized algorithm to solve the problem, which also offers approximation guarantees. Finally, we demonstrate that, by incorporating a matroid constraint into the problem formulation, we can increase the diversity of the optimal set of counterfactual explanations and incentivize individuals across the whole spectrum of the population to self improve. Experiments on synthetic and real lending and credit card data illustrate our theoretical findings and show that the counterfactual explanations and decision policies found by our algorithms achieve higher utility than several competitive baselines.

[Hierarchical Patch VAE-GAN: Generating Diverse Videos from a Single Sample](#)

- Shir Gur, Sagie Benaim, Lior Wolf
- abstract@[open-review](#): We consider the task of generating diverse and novel videos from a single video sample. Recently, new hierarchical patch-GAN based approaches were proposed for generating diverse images, given only a single sample at training time. Moving to videos, these approaches fail to generate diverse samples, and often collapse into generating samples similar to the training video. We introduce a novel patch-based variational autoencoder (VAE) which allows for a much greater diversity in generation. Using this tool, a new hierarchical video generation scheme is constructed: at coarse scales, our patch-VAE is employed, ensuring samples are of high diversity. Subsequently, at finer scales, a patch-GAN renders the fine details, resulting in high quality videos. Our experiments show that the proposed method produces diverse samples in both the image domain, and the more challenging video domain. Our code and supplementary material (SM) with additional samples are available at <https://shirgur.github.io/hp-vae-gan>

[A Feasible Level Proximal Point Method for Nonconvex Sparse Constrained Optimization](#)

- Digvijay Boob, Qi Deng, Guanghui Lan, Yilin Wang
- abstract@[open-review](#): Nonconvex sparse models have received significant attention in high-dimensional machine learning. In this paper, we study a new model consisting of a general convex or nonconvex objectives and a variety of continuous nonconvex sparsity-inducing constraints. For this constrained model, we propose a novel proximal point algorithm that solves a sequence of convex subproblems with gradually relaxed constraint levels. Each subproblem, having a proximal point objective and a convex surrogate constraint, can be efficiently solved based on a fast routine for projection onto the surrogate constraint. We establish the asymptotic convergence of the proposed algorithm to the Karush-Kuhn-Tucker (KKT) solutions. We also establish new convergence complexities to achieve an approximate KKT solution when the objective can be smooth/nonsmooth, deterministic/stochastic and convex/nonconvex with complexity that is on a par with gradient descent for unconstrained optimization problems in respective cases. To the best of our knowledge, this is the first study of the first-order methods with complexity guarantee for nonconvex sparse-constrained problems. We perform numerical experiments to demonstrate the effectiveness of our new model and efficiency of the proposed algorithm for large scale problems.

[Reservoir Computing meets Recurrent Kernels and Structured Transforms](#)

- Jonathan Dong, Ruben Ohana, Mushegh Rafayelyan, Florent Krzakala
- abstract@[open-review](#): Reservoir Computing is a class of simple yet efficient Recurrent Neural Networks where internal weights are fixed at random and only a linear output layer is trained. In the large size limit, such random neural networks have a deep connection with kernel methods. Our contributions are threefold: a) We rigorously establish the recurrent kernel limit of Reservoir Computing and prove its convergence. b) We test our models on chaotic time series prediction, a classic but challenging benchmark in Reservoir Computing, and show how the Recurrent Kernel is competitive and computationally efficient when the number of data points remains moderate. c) When the number of samples is too large, we leverage the success of structured Random Features for kernel approximation by introducing Structured Reservoir Computing. The two proposed methods, Recurrent Kernel and Structured Reservoir Computing, turn out to be much faster and more memory-efficient than conventional Reservoir Computing.

[Comprehensive Attention Self-Distillation for Weakly-Supervised Object Detection](#)

- Zeyi Huang, Yang Zou, B. V. K. Vijaya Kumar, Dong Huang
- abstract@[open-review](#): Weakly Supervised Object Detection (WSOD) has emerged as an effective tool to train object detectors using only the image-level category labels. However, without object-level labels, WSOD detectors are prone to detect bounding boxes on salient objects, clustered objects and discriminative object parts. Moreover, the image-level category labels do not enforce consistent object detection across different transformations of the same images. To address the above issues, we propose a Comprehensive Attention Self-Distillation (CASD) training approach for WSOD. To balance feature learning among all object instances, CASD computes the comprehensive attention aggregated from multiple transformations and feature layers of the same images. To enforce consistent spatial supervision on objects, CASD conducts self-distillation on the WSOD networks, such that the comprehensive attention is approximated simultaneously by multiple transformations and feature layers of the same images. CASD produces new state-of-the-art WSOD results on standard benchmarks such as PASCAL VOC 2007/2012 and MS-COCO.

[Linear Dynamical Systems as a Core Computational Primitive](#)

- Shiva Kaul
- abstract@[open-review](#): Running nonlinear RNNs for T steps takes $O(T)$ time. Our construction, called LDStack, approximately runs them in $O(\log T)$ parallel time, and obtains arbitrarily low error via repetition. First, we show nonlinear RNNs can be approximated by a stack of multiple-input, multiple-output (MIMO) LDS. This replaces nonlinearity across time with nonlinearity along depth. Next, we show that MIMO LDS can be approximated by an average or a concatenation of single-input, multiple-output (SIMO) LDS. Finally, we present an algorithm for running (and differentiating) SIMO LDS in $O(\log T)$ parallel time. On long sequences, LDStack is much faster than traditional RNNs, yet it achieves similar accuracy in our experiments. Furthermore, LDStack is amenable to linear systems theory. Therefore, it improves not only speed, but also interpretability and mathematical tractability.

[Ratio Trace Formulation of Wasserstein Discriminant Analysis](#)

- Hexuan Liu, Yunfeng Cai, You-Lin Chen, Ping Li
- abstract@[open-review](#): We reformulate the Wasserstein Discriminant Analysis (WDA) as a ratio trace problem and present an eigensolver-based algorithm to compute the discriminative subspace of WDA. This new formulation, along with the proposed algorithm, can be served as an efficient and more stable alternative to the original trace ratio formulation and its gradient-based algorithm. We provide a rigorous convergence analysis for the proposed algorithm under the self-consistent field framework, which is crucial but missing in the literature. As an application, we combine WDA with low-dimensional clustering techniques, such as K-means, to perform subspace clustering. Numerical experiments on real datasets show promising results of the ratio trace formulation of WDA in both classification and clustering tasks.

[PAC-Bayes Analysis Beyond the Usual Bounds](#)

- Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvari, John Shawe-Taylor
- abstract@[open-review](#): We focus on a stochastic learning model where the learner observes a finite set of training examples and the output of the learning process is a data-dependent distribution over a space of hypotheses. The learned data-dependent distribution is then used to make randomized predictions, and the high-level theme addressed here is guaranteeing the quality of predictions on examples that were not seen during training, i.e. generalization. In this setting the unknown quantity of interest is the expected risk of the data-dependent randomized predictor, for which upper bounds can be derived via a PAC-Bayes analysis, leading to PAC-Bayes bounds. Specifically, we present a basic PAC-Bayes inequality for stochastic kernels, from which one may derive extensions of various known PAC-Bayes bounds as well as novel bounds. We clarify the role of the requirements of fixed “data-free” priors, bounded losses, and i.i.d. data. We highlight that those requirements were used to upper-bound an exponential moment term, while the basic PAC-Bayes theorem remains valid without those restrictions. We present three bounds that illustrate the use of data-dependent priors, including one for the unbounded square loss.

[Few-shot Visual Reasoning with Meta-Analogical Contrastive Learning](#)

- Youngsung Kim, Jinwoo Shin, Eunho Yang, Sung Ju Hwang
- abstract@[open-review](#): While humans can solve a visual puzzle that requires logical reasoning by observing only few samples, it would require training over a large number of samples for state-of-the-art deep reasoning models to obtain similar performance on the same task. In this work, we propose to solve such a few-shot (or low-shot) abstract visual reasoning problem by resorting to \emph{analogical reasoning}, which is a unique human ability to identify structural or relational similarity between two sets. Specifically, we construct analogical and non-analogical training pairs of two different problem instances, e.g., the latter is created by perturbing or shuffling the original (former) problem. Then, we extract the structural relations among elements in both domains in a pair by enforcing analogical ones to be as similar as possible, while minimizing similarities between non-analogical ones. This analogical contrastive learning allows to effectively learn the relational representations of given abstract reasoning tasks. We validate our method on RAVEN dataset, on which it outperforms state-of-the-art method, with larger gains when the training data is scarce. We further meta-learn our analogical contrastive learning model over the same tasks with diverse attributes, and show that it generalizes to the same visual reasoning problem with unseen attributes.

[MPNet: Masked and Permuted Pre-training for Language Understanding](#)

- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu
- abstract@[open-review](#): BERT adopts masked language modeling (MLM) for pre-training and is one of the most successful pre-training models. Since BERT neglects dependency among predicted tokens, XLNet introduces permuted language modeling (PLM) for pre-training to address this problem. However, XLNet does not leverage the full position information of a sentence and thus suffers from position discrepancy between pre-training and fine-tuning. In this paper, we propose MPNet, a novel pre-training method that inherits the advantages of BERT and XLNet and avoids their limitations. MPNet leverages the dependency among predicted tokens through permuted language modeling (vs. MLM in BERT), and takes auxiliary position information as input to make the model see a full sentence and thus reducing the position discrepancy (vs. PLM in XLNet). We pre-train MPNet on a large-scale dataset (over 160GB text corpora) and fine-tune on a variety of down-streaming tasks (GLUE, SQuAD, etc). Experimental results show that MPNet outperforms MLM and PLM by a large margin, and achieves better results on these tasks compared with previous state-of-the-art pre-trained methods (e.g., BERT, XLNet, RoBERTa) under the same model setting. We attach the code in the supplemental materials.

[Reinforcement Learning with Feedback Graphs](#)

- Christoph Dann, Yishay Mansour, Mehryar Mohri, Ayush Sekhari, Karthik Sridharan
- abstract@[open-review](#): We study RL in the tabular MDP setting where the agent receives additional observations per step in the form of transitions samples. Such additional observations can be provided in many tasks by auxiliary sensors or by leveraging prior knowledge about the environment (e.g., when certain actions yield similar outcome). We formalize this setting using a feedback graph over state-action pairs and show that model-based algorithms can incorporate additional observations for more sample-efficient learning. We give a regret bound that predominantly depends on the size of the maximum acyclic subgraph of the feedback graph, in contrast with a polynomial dependency on the number of states and actions in the absence of side observations. Finally, we highlight fundamental challenges for leveraging a small dominating set of the feedback graph, as compared to the well-studied bandit setting, and propose a new algorithm that can use such a dominating set to learn a near-optimal policy faster.

[Zap Q-Learning With Nonlinear Function Approximation](#)

- Shuhang Chen, Adithya M Devraj, Fan Lu, Ana Busic, Sean Meyn
- abstract@[open-review](#): Zap Q-learning is a recent class of reinforcement learning algorithms, motivated primarily as a means to accelerate convergence. Stability theory has been absent outside of two restrictive classes: the tabular setting, and optimal stopping. This paper introduces a new framework for analysis of a more general class of recursive algorithms known as stochastic approximation. Based on this general theory, it is shown that Zap Q-learning is consistent under a non-degeneracy assumption, even when the function approximation architecture is nonlinear. Zap Q-learning with neural network function approximation emerges as a special case, and is tested on examples from OpenAI Gym. Based on multiple experiments with a range of neural network sizes, it is found that the new algorithms converge quickly and are robust to choice of function approximation architecture.

[Lipschitz-Certifiable Training with a Tight Outer Bound](#)

- Sungyoon Lee, Jaewook Lee, Saerom Park
- abstract@[open-review](#): Verifiable training is a promising research direction for training a robust network. However, most verifiable training methods are slow or lack scalability. In this study, we propose a fast and scalable certifiable training algorithm based on Lipschitz analysis and interval arithmetic. Our certifiable training algorithm provides a tight propagated outer bound by introducing the box constraint propagation (BCP), and it efficiently computes the worst logit over the outer bound. In the experiments, we show that BCP achieves a tighter outer bound than the global Lipschitz-based outer bound. Moreover, our certifiable training algorithm is over 12 times faster than the state-of-the-art dual relaxation-based method; however, it achieves comparable or better verification performance, improving natural accuracy. Our fast certifiable training algorithm with the tight outer bound can scale to Tiny ImageNet with verification accuracy of 20.1% (ℓ_2 -perturbation of $\epsilon=36/255$). Our code is available at [url{https://github.com/sungyoon-lee/bcp}](https://github.com/sungyoon-lee/bcp).

[Fast Adaptive Non-Monotone Submodular Maximization Subject to a Knapsack Constraint](#)

- Georgios Amanatidis, Federico Fusco, Philip Lazos, Stefano Leonardi, Rebecca Reiffen
- abstract@[open-review](#): Constrained submodular maximization problems encompass a wide variety of applications, including personalized recommendation, team formation, and revenue maximization via viral marketing. The massive instances occurring in modern-day applications can render existing algorithms prohibitively slow. Moreover, frequently those instances are also inherently stochastic. Focusing on these challenges, we revisit the classic problem of maximizing a (possibly non-monotone) submodular function subject to a knapsack constraint. We present a simple randomized greedy algorithm that achieves a $\$5.83\$$ approximation and runs in $\$O(n \log n)$ time, i.e., at least a factor n faster than other state-of-the-art algorithms. The robustness of our approach allows us to further transfer it to a stochastic version of the problem. There, we obtain a 9-approximation to the best adaptive policy, which is the first constant approximation for non-monotone objectives. Experimental evaluation of our algorithms showcases their improved performance on real and synthetic data.

[Conformal Symplectic and Relativistic Optimization](#)

- Guilherme Franca, Jeremias Sulam, Daniel Robinson, Rene Vidal
- abstract@[open-review](#): Arguably, the two most popular accelerated or momentum-based optimization methods are Nesterov's accelerated gradient and Polyak's heavy ball, both corresponding to different discretizations of a particular second order differential equation with a friction term. Such connections with continuous-time dynamical systems have been instrumental in demystifying acceleration phenomena in optimization. Here we study structure-preserving discretizations for a certain class of dissipative (conformal) Hamiltonian systems, allowing us to analyze the symplectic structure of both Nesterov and heavy ball, besides providing several new insights into these methods. Moreover, we propose a new algorithm based on a dissipative relativistic system that normalizes the momentum and may result in more stable/faster optimization. Importantly, such a method generalizes both Nesterov and heavy ball, each being recovered as distinct limiting cases, and has potential advantages at no additional cost.

[Bayes Consistency vs. H-Consistency: The Interplay between Surrogate Loss Functions and the Scoring Function Class](#)

- Mingyuan Zhang, Shivani Agarwal
- abstract@[open-review](#): A fundamental question in multiclass classification concerns understanding the consistency properties of surrogate risk minimization algorithms, which minimize a (often convex) surrogate to the multiclass 0-1 loss. In particular, the framework of calibrated surrogates has played an important role in analyzing the Bayes consistency properties of such algorithms, i.e. in studying convergence to a Bayes optimal classifier (Zhang, 2004; Tewari and Bartlett, 2007). However, follow-up work has suggested this framework can be of limited value when studying H-consistency; in particular, concerns have been raised that even when the data comes from an underlying linear model, minimizing certain convex calibrated surrogates over linear scoring functions fails to recover the true model (Long and Servedio, 2013). In this paper, we investigate this apparent conundrum. We find that while some calibrated surrogates can indeed fail to provide H-consistency when minimized over a natural-looking but naively chosen scoring function class F , the situation can potentially be remedied by minimizing them over a more carefully chosen class of scoring functions F . In particular, for the popular one-vs-all hinge and logistic surrogates, both of which are calibrated (and therefore provide Bayes consistency) under realizable models, but were previously shown to pose problems for realizable H-consistency, we derive a form of scoring function class F that enables H-consistency. When H is the class of linear models, the class F consists of certain piecewise linear scoring functions that are characterized by the same number of parameters as in the linear case, and minimization over which can be performed using an adaptation of the min-pooling idea from neural network training. Our experiments confirm that the one-vs-all surrogates, when trained over this class of nonlinear scoring functions F , yield better linear multiclass classifiers than when trained over standard linear scoring functions.

[Inverting Gradients - How easy is it to break privacy in federated learning?](#)

- Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, Michael Moeller
- abstract@[open-review](#): The idea of federated learning is to collaboratively train a neural network on a server. Each user receives the current weights of the network and in turns sends parameter updates (gradients) based on local data. This protocol has been designed not only to train neural networks data-efficiently, but also to provide privacy benefits for users, as their input data remains on device and only parameter gradients are shared. But how secure is sharing parameter gradients? Previous attacks have provided a false sense of security, by succeeding only in contrived settings - even for a single image. However, by exploiting a magnitude-invariant loss along with optimization strategies based on adversarial attacks, we show that it is actually possible to faithfully reconstruct images at high resolution from the knowledge of their parameter gradients, and demonstrate that such a break of privacy is possible even for trained deep networks. We analyze the effects of architecture as well as parameters on the difficulty of reconstructing an input image and prove that any input to a fully connected layer can be reconstructed analytically independent of the remaining architecture. Finally we discuss settings encountered in practice and show that even averaging gradients over several iterations or several images does not protect the user's privacy in federated learning applications.

[Dynamic allocation of limited memory resources in reinforcement learning](#)

- Nisheet Patel, Luigi Acerbi, Alexandre Pouget
- abstract@[open-review](#): Biological brains are inherently limited in their capacity to process and store information, but are nevertheless capable of solving complex tasks with apparent ease. Intelligent behavior is related to these limitations, since resource constraints drive the need to generalize and assign importance differentially to features in the environment or memories of past experiences. Recently, there have been parallel efforts in reinforcement learning and neuroscience to understand strategies adopted by artificial and biological agents to circumvent limitations in information storage. However, the two threads have been largely separate. In this article, we propose a dynamical framework to maximize expected reward under constraints of limited resources, which we implement with a cost function that penalizes precise representations of action-values in memory, each of which may vary in its precision. We derive from first principles an algorithm, Dynamic Resource Allocator (DRA), which we apply to two standard tasks in reinforcement learning and a model-based planning task, and find that it allocates more resources to items in memory that have a higher impact on cumulative rewards. Moreover, DRA learns faster when starting with a higher resource budget than what it eventually allocates for performing well on tasks, which may explain why frontal cortical areas in biological brains appear more engaged in early stages of learning before settling to lower asymptotic levels of activity. Our work provides a normative solution to the problem of learning how to allocate costly resources to a collection of uncertain memories in a manner that is capable of adapting to changes in the environment.

[CryptoNAS: Private Inference on a ReLU Budget](#)

- Zahra Ghodsi, Akshaj Kumar Veldanda, Brandon Reagen, Siddharth Garg
- abstract@[open-review](#): Machine learning as a service has given raise to privacy concerns surrounding clients' data and providers' models and has catalyzed research in private inference (PI): methods to process inferences without disclosing inputs. Recently, researchers have adapted cryptographic techniques to show PI is possible, however all solutions increase inference latency beyond practical limits. This paper makes the observation that existing models are ill-suited for PI and proposes a novel NAS method, named CryptoNAS, for finding and tailoring models to the needs of PI. The key insight is that in PI operator latency cost are inverted: non-linear operations (e.g., ReLU) dominate latency, while linear layers become effectively free. We develop the idea of a ReLU budget as a proxy for inference latency and use CryptoNAS to build models that maximize accuracy within a given budget. CryptoNAS improves accuracy by 3.4% and latency by 2.4x over the state-of-the-art.

[A Stochastic Path Integral Differential Estimator Expectation Maximization Algorithm](#)

- Gersende Fort, Eric Moulines, Hoi-To Wai
- abstract@[open-review](#): The Expectation Maximization (EM) algorithm is of key importance for inference in latent variable models including mixture of regressors and experts, missing observations. This paper introduces a novel EM algorithm, called {\tt SPIDER-EM}, for inference from a training set of size \$n\$, \$n \gg 1\$. At the core of our algorithm is an estimator of the full conditional expectation in the {\sf E}-step, adapted from the stochastic path integral differential estimator ({\tt SPIDER}) technique. We derive finite-time complexity bounds for smooth non-convex likelihood: we show that for convergence to an \$\epsilon\$-approximate stationary point, the complexity scales as \$K_{\text{Opt}}(n, \epsilon) = \mathcal{O}(\epsilon^{-1})\$ and \$K_{\text{CE}}(n, \epsilon) = n + \sqrt{n} \mathcal{O}(\epsilon^{-1})\$, where \$K_{\text{Opt}}(n, \epsilon)\$ and \$K_{\text{CE}}(n, \epsilon)\$ are respectively the number of {\sf M}-steps and the number of per-sample conditional expectations evaluations. This improves over the state-of-the-art algorithms. Numerical results support our findings.

[CHIP: A Hawkes Process Model for Continuous-time Networks with Scalable and Consistent Estimation](#)

- Makan Arastuie, Subhadeep Paul, Kevin Xu
- abstract@[open-review](#): In many application settings involving networks, such as messages between users of an on-line social network or transactions between traders in financial markets, the observed data consist of timestamped relational events, which form a continuous-time network. We propose the Community Hawkes Independent Pairs (CHIP) generative model for such networks. We show that applying spectral clustering to an aggregated adjacency matrix constructed from the CHIP model provides consistent community detection for a growing number of nodes and time duration. We also develop consistent and computationally efficient estimators for the model parameters. We demonstrate that our proposed CHIP model and estimation procedure scales to large networks with tens of thousands of nodes and provides superior fits than existing continuous-time network models on several real networks.

[SAC: Accelerating and Structuring Self-Attention via Sparse Adaptive Connection](#)

- Xiaoya Li, Yuxian Meng, Mingxin Zhou, Qinghong Han, Fei Wu, Jiwei Li
- abstract@[open-review](#): While the self-attention mechanism has been widely used in a wide variety of tasks, it has the unfortunate property of a quadratic cost with respect to the input length, which makes it difficult to deal with long inputs. In this paper, we present a method for accelerating and structuring self-attentions: Sparse Adaptive Connection (SAC). In SAC, we regard the input sequence as a graph and attention operations are performed between linked nodes. In contrast with previous self-attention models with pre-defined structures (edges), the model learns to construct attention edges to improve task-specific performances. In this way, the model is able to select the most salient nodes and reduce the quadratic complexity regardless of the sequence length. Based on SAC, we show that previous variants of self-attention models are its special cases. Through extensive experiments on neural machine translation, language modeling, graph representation learning and image classification, we demonstrate SAC is competitive with state-of-the-art models while significantly reducing memory cost.

[Design Space for Graph Neural Networks](#)

- Jiaxuan You, Zhitao Ying, Jure Leskovec
- abstract@[open-review](#): The rapid evolution of Graph Neural Networks (GNNs) has led to a growing number of new architectures as well as novel applications. However, current research focuses on proposing and evaluating specific architectural designs of GNNs, such as GCN, GIN, or GAT, as opposed to studying the more general design space of GNNs that consists of a Cartesian product of different design dimensions, such as the number of layers or the type of the aggregation function. Additionally, GNN designs are often specialized to a single task, yet few efforts have been made to understand how to quickly find the best GNN design for a novel task or a novel dataset. Here we define and systematically study the architectural design space for GNNs which consists of 315,000 different designs over 32 different predictive tasks. Our approach features three key innovations: (1) A general GNN design space; (2) a GNN task space with a similarity metric, so that for a given novel task/dataset, we can quickly identify/transfer the best performing architecture; (3) an efficient and effective design space evaluation method which allows insights to be distilled from a huge number of model-task combinations. Our key results include: (1) A comprehensive set of guidelines for designing well-performing GNNs; (2) while best GNN designs for different tasks vary significantly, the GNN task space allows for transferring the best designs across different tasks; (3) models discovered using our design space achieve state-of-the-art performance. Overall, our work offers a principled and scalable approach to transition from studying individual GNN designs for specific tasks, to systematically studying the GNN design space and the task space. Finally, we release GraphGym, a powerful platform for exploring different GNN designs and tasks. GraphGym features modularized GNN implementation, standardized GNN evaluation, and reproducible and scalable experiment management.

[HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis](#)

- Jungil Kong, Jaehyeon Kim, Jaekyoung Bae
- abstract@[open-review](#): Several recent work on speech synthesis have employed generative adversarial networks (GANs) to produce raw waveforms. Although such methods improve the sampling efficiency and memory usage, their sample quality has not yet reached that of autoregressive and flow-based generative models. In this work, we propose HiFi-GAN, which achieves both efficient and high-fidelity speech synthesis. As speech audio consists of sinusoidal signals with various periods, we demonstrate that modeling periodic patterns of an audio is crucial for enhancing sample quality. A subjective human evaluation (mean opinion score, MOS) of a single speaker dataset indicates that our proposed method demonstrates similarity to human quality while generating 22.05 kHz high-fidelity audio 167.9 times faster than real-time on a single V100 GPU. We further show the generality of HiFi-GAN to the mel-spectrogram inversion of unseen speakers and end-to-end speech synthesis. Finally, a small footprint version of HiFi-GAN generates samples 13.4 times faster than real-time on CPU with comparable quality to an autoregressive counterpart.

[Unbalanced Sobolev Descent](#)

- Youssef Mroueh, Mattia Rigotti
- abstract@[open-review](#): We introduce Unbalanced Sobolev Descent (USD), a particle descent algorithm for transporting a high dimensional source distribution to a target distribution that does not necessarily have the same mass. We define the Sobolev-Fisher discrepancy between distributions and show that it relates to advection-reaction transport equations and the Wasserstein-Fisher-Rao metric between distributions. USD transports particles along gradient flows of the witness function of the Sobolev-Fisher discrepancy (advection step) and reweights the mass of particles with respect to this witness function (reaction step). The reaction step can be thought of as a birth-death process of the particles with rate of growth proportional to the witness function. When the Sobolev-Fisher witness function is estimated in a Reproducing Kernel Hilbert Space (RKHS), under mild assumptions we show that

USD converges asymptotically (in the limit of infinite particles) to the target distribution in the Maximum Mean Discrepancy (MMD) sense. We then give two methods to estimate the Sobolev-Fisher witness with neural networks, resulting in two Neural USD algorithms. The first one implements the reaction step with mirror descent on the weights, while the second implements it through a birth-death process of particles. We show on synthetic examples that USD transports distributions with or without conservation of mass faster than previous particle descent algorithms, and finally demonstrate its use for molecular biology analyses where our method is naturally suited to match developmental stages of populations of differentiating cells based on their single-cell RNA sequencing profile. Code is available at <http://github.com/ibm/usd>.

[Identifying Mislabeled Data using the Area Under the Margin Ranking](#)

- Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, Kilian Q. Weinberger
- abstract@[open-review](#): Not all data in a typical training set help with generalization; some samples can be overly ambiguous or outrightly mislabeled. This paper introduces a new method to identify such samples and mitigate their impact when training neural networks. At the heart of our algorithm is the Area Under the Margin (AUM) statistic, which exploits differences in the training dynamics of clean and mislabeled samples. A simple procedure - adding an extra class populated with purposefully mislabeled threshold samples - learns a AUM upper bound that isolates mislabeled data. This approach consistently improves upon prior work on synthetic and real-world datasets. On the WebVision50 classification task our method removes 17% of training data, yielding a 1.6% (absolute) improvement in test error. On CIFAR100 removing 13% of the data leads to a 1.2% drop in error.

[Combining Deep Reinforcement Learning and Search for Imperfect-Information Games](#)

- Noam Brown, Anton Bakhtin, Adam Lerer, Qucheng Gong
- abstract@[open-review](#): The combination of deep reinforcement learning and search at both training and test time is a powerful paradigm that has led to a number of successes in single-agent settings and perfect-information games, best exemplified by AlphaZero. However, prior algorithms of this form cannot cope with imperfect-information games. This paper presents ReBeL, a general framework for self-play reinforcement learning and search that provably converges to a Nash equilibrium in any two-player zero-sum game. In the simpler setting of perfect-information games, ReBeL reduces to an algorithm similar to AlphaZero. Results in two different imperfect-information games show ReBeL converges to an approximate Nash equilibrium. We also show ReBeL achieves superhuman performance in heads-up no-limit Texas hold'em poker, while using far less domain knowledge than any prior poker AI.

[High-Throughput Synchronous Deep RL](#)

- Iou-Jen Liu, Raymond Yeh, Alexander Schwing
- abstract@[open-review](#): Various parallel actor-learner methods reduce long training times for deep reinforcement learning. Synchronous methods enjoy training stability while having lower data throughput. In contrast, asynchronous methods achieve high throughput but suffer from stability issues and lower sample efficiency due to ∞ -stale policies.TM To combine the advantages of both methods we propose High-Throughput Synchronous Deep Reinforcement Learning (HTS-RL). In HTS-RL, we perform learning and rollouts concurrently, devise a system design which avoids ∞ -stale policiesTM and ensure that actors interact with environment replicas in an asynchronous manner while maintaining full determinism. We evaluate our approach on Atari games and the Google Research Football environment. Compared to synchronous baselines, HTS-RL is 2 \sim 6X faster. Compared to state-of-the-art asynchronous methods, HTS-RL has competitive throughput and consistently achieves higher average episode rewards.

[Contrastive Learning with Adversarial Examples](#)

- Chih-Hui Ho, Nuno Vasconcelos
- abstract@[open-review](#): Contrastive learning (CL) is a popular technique for self-supervised learning (SSL) of visual representations. It uses pairs of augmentations of unlabeled training examples to define a classification task for pretext learning of a deep embedding. Despite extensive works in augmentation procedures, prior works do not address the selection of challenging negative pairs, as images within a sampled batch are treated independently. This paper addresses the problem, by introducing a new family of adversarial examples for contrastive learning and using these examples to define a new adversarial training algorithm for SSL, denoted as CLAE. When compared to standard CL, the use of adversarial examples creates more challenging positive pairs and adversarial training produces harder negative pairs by accounting for all images in a batch during the optimization. CLAE is compatible with many CL methods in the literature. Experiments show that it improves the performance of several existing CL baselines on multiple datasets.

[Mixed Hamiltonian Monte Carlo for Mixed Discrete and Continuous Variables](#)

- Guangyao Zhou
- abstract@[open-review](#): Hamiltonian Monte Carlo (HMC) has emerged as a powerful Markov Chain Monte Carlo (MCMC) method to sample from complex continuous distributions. However, a fundamental limitation of HMC is that it can not be applied to distributions with mixed discrete and continuous variables. In this paper, we propose mixed HMC (M-HMC) as a general framework to address this limitation. M-HMC is a novel family of MCMC algorithms that evolves the discrete and continuous variables in tandem, allowing more frequent updates of discrete variables while maintaining HMC's ability to suppress random-walk behavior. We establish M-HMC's theoretical properties, and present an efficient implementation with Laplace momentum that introduces minimal overhead compared to existing HMC methods. The superior performances of M-HMC over existing methods are demonstrated with numerical experiments on Gaussian mixture models (GMMs), variable selection in Bayesian logistic regression (BLR), and correlated topic models (CTMs).

[Adversarial Sparse Transformer for Time Series Forecasting](#)

- Sifan Wu, Xi Xiao, Qianggang Ding, Peilin Zhao, Ying Wei, Junzhou Huang
- abstract@[open-review](#): Many approaches have been proposed for time series forecasting, in light of its significance in wide applications including business demand prediction. However, the existing methods suffer from two key limitations. Firstly, most point prediction models only predict an exact value of each time step without flexibility, which can hardly capture the stochasticity of data. Even probabilistic prediction using the likelihood estimation suffers these problems in the same way. Besides, most of them use the auto-regressive generative mode, where ground-truth is provided during training and replaced by the network's own one-step ahead output during inference, causing the error accumulation in inference. Thus they may fail to forecast time series for long time horizon due to the error accumulation. To solve these issues, in this paper, we propose a new time series forecasting model -- Adversarial Sparse Transformer (AST), based on Generated Adversarial Networks (GANs). Specifically, AST adopts a Sparse Transformer as the generator to learn a sparse attention map for time series forecasting, and uses a discriminator to improve the prediction performance from sequence level. Extensive experiments on several real-world datasets show the effectiveness and efficiency of our method.

[The Surprising Simplicity of the Early-Time Learning Dynamics of Neural Networks](#)

- Wei Hu, Lechao Xiao, Ben Adlam, Jeffrey Pennington
- abstract@[open-review](#): Modern neural networks are often regarded as complex black-box functions whose behavior is difficult to understand owing to their nonlinear dependence on the data and the nonconvexity in their loss landscapes. In this work, we show that these common perceptions can be

completely false in the early phase of learning. In particular, we formally prove that, for a class of well-behaved input distributions, the early-time learning dynamics of a two-layer fully-connected neural network can be mimicked by training a simple linear model on the inputs. We additionally argue that this surprising simplicity can persist in networks with more layers and with convolutional architecture, which we verify empirically. Key to our analysis is to bound the spectral norm of the difference between the Neural Tangent Kernel (NTK) and an affine transform of the data kernel; however, unlike many previous results utilizing the NTK, we do not require the network to have disproportionately large width, and the network is allowed to escape the kernel regime later in training.

[CLEARER: Multi-Scale Neural Architecture Search for Image Restoration](#)

- Yuanbiao Gou, Boyun Li, Zitao Liu, Songfan Yang, Xi Peng
- abstract@[open-review](#): Multi-scale neural networks have shown effectiveness in image restoration tasks, which are usually designed and integrated in a handcrafted manner. Different from the existing labor-intensive handcrafted architecture design paradigms, we present a novel method, termed as multi-sCaLe nEral ARchitecture sEarch for image Restoration (CLEARER), which is a specifically designed neural architecture search (NAS) for image restoration. Our contributions are twofold. On one hand, we design a multi-scale search space that consists of three task-exible modules. Namely, 1) Parallel module that connects multi-resolution neural blocks in parallel, while preserving the channels and spatial-resolution in each neural block, 2) Transition module remains the existing multi-resolution features while extending them to a lower resolution, 3) Fusion module integrates multi-resolution features by passing the features of the parallel neural blocks to the current neural blocks. On the other hand, we present novel losses which could 1) balance the tradeoff between the model complexity and performance, which is highly expected to image restoration; and 2) relax the discrete architecture parameters into a continuous distribution which approximates to either 0 or 1. As a result, a differentiable strategy could be employed to search when to fuse or extract multi-resolution features, while the discretization issue faced by the gradient-based NAS could be alleviated. The proposed CLEARER could search a promising architecture in two GPU hours. Extensive experiments show the promising performance of our method comparing with nine image denoising methods and eight image deraining approaches in quantitative and qualitative evaluations. The codes are available at <https://github.com/limit-scu>.

[Hierarchical Gaussian Process Priors for Bayesian Neural Network Weights](#)

- Theofanis Karaletsos, Thang D. Bui
- abstract@[open-review](#): Probabilistic neural networks are typically modeled with independent weight priors, which do not capture weight correlations in the prior and do not provide a parsimonious interface to express properties in function space. A desirable class of priors would represent weights compactly, capture correlations between weights, facilitate calibrated reasoning about uncertainty, and allow inclusion of prior knowledge about the function space such as periodicity or dependence on contexts such as inputs. To this end, this paper introduces two innovations: (i) a Gaussian process-based hierarchical model for network weights based on unit embeddings that can flexibly encode correlated weight structures, and (ii) input-dependent versions of these weight priors that can provide convenient ways to regularize the function space through the use of kernels defined on contextual inputs. We show these models provide desirable test-time uncertainty estimates on out-of-distribution data, demonstrate cases of modeling inductive biases for neural networks with kernels which help both interpolation and extrapolation from training data, and demonstrate competitive predictive performance on an active learning benchmark.

[Compositional Explanations of Neurons](#)

- Jesse Mu, Jacob Andreas
- abstract@[open-review](#): We describe a procedure for explaining neurons in deep representations by identifying compositional logical concepts that closely approximate neuron behavior. Compared to prior work that uses atomic labels as explanations, analyzing neurons compositionally allows us to more precisely and expressively characterize their behavior. We use this procedure to answer several questions on interpretability in models for vision and natural language processing. First, we examine the kinds of abstractions learned by neurons. In image classification, we find that many neurons learn highly abstract but semantically coherent visual concepts, while other polysemantic neurons detect multiple unrelated features; in natural language inference (NLI), neurons learn shallow lexical heuristics from dataset biases. Second, we see whether compositional explanations give us insight into model performance: vision neurons that detect human-interpretable concepts are positively correlated with task performance, while NLI neurons that fire for shallow heuristics are negatively correlated with task performance. Finally, we show how compositional explanations provide an accessible way for end users to produce simple "copy-paste" adversarial examples that change model behavior in predictable ways.

[Calibrated Reliable Regression using Maximum Mean Discrepancy](#)

- Peng Cui, Wenbo Hu, Jun Zhu
- abstract@[open-review](#): Accurate quantification of uncertainty is crucial for real-world applications of machine learning. However, modern deep neural networks still produce unreliable predictive uncertainty, often yielding over-confident predictions. In this paper, we are concerned with getting well-calibrated predictions in regression tasks. We propose the calibrated regression method using the maximum mean discrepancy by minimizing the kernel embedding measure. Theoretically, the calibration error of our method asymptotically converges to zero when the sample size is large enough. Experiments on non-trivial real datasets show that our method can produce well-calibrated and sharp prediction intervals, which outperforms the related state-of-the-art methods.

[Directional convergence and alignment in deep learning](#)

- Ziwei Ji, Matus Telgarsky
- abstract@[open-review](#): In this paper, we show that although the minimizers of cross-entropy and related classification losses are off at infinity, network weights learned by gradient flow converge in direction, with an immediate corollary that network predictions, training errors, and the margin distribution also converge. This proof holds for deep homogeneous networks – a broad class of networks allowing for ReLU, max-pooling, linear, and convolutional layers – and we additionally provide empirical support not just close to the theory (e.g., the AlexNet), but also on non-homogeneous networks (e.g., the DenseNet). If the network further has locally Lipschitz gradients, we show that these gradients also converge in direction, and asymptotically align with the gradient flow path, with consequences on margin maximization, convergence of saliency maps, and a few other settings. Our analysis complements and is distinct from the well-known neural tangent and mean-field theories, and in particular makes no requirements on network width and initialization, instead merely requiring perfect classification accuracy. The proof proceeds by developing a theory of unbounded nonsmooth Kurdyka-Å ojasiewicz inequalities for functions definable in an o-minimal structure, and is also applicable outside deep learning.

[Functional Regularization for Representation Learning: A Unified Theoretical Perspective](#)

- Siddhant Garg, Yingyu Liang
- abstract@[open-review](#): Unsupervised and self-supervised learning approaches have become a crucial tool to learn representations for downstream prediction tasks. While these approaches are widely used in practice and achieve impressive empirical gains, their theoretical understanding largely lags behind. Towards bridging this gap, we present a unifying perspective where several such approaches can be viewed as imposing a regularization on the representation via a learnable function using unlabeled data. We propose a discriminative theoretical framework for analyzing the sample complexity of these approaches, which generalizes the framework of (Balcan and Blum, 2010) to allow learnable regularization functions. Our sample complexity

bounds show that, with carefully chosen hypothesis classes to exploit the structure in the data, these learnable regularization functions can prune the hypothesis space, and help reduce the amount of labeled data needed. We then provide two concrete examples of functional regularization, one using auto-encoders and the other using masked self-supervision, and apply our framework to quantify the reduction in the sample complexity bound of labeled data. We also provide complementary empirical results to support our analysis.

[Provably Efficient Online Hyperparameter Optimization with Population-Based Bandits](#)

- Jack Parker-Holder, Vu Nguyen, Stephen J. Roberts
- abstract@[open-review](#): Many of the recent triumphs in machine learning are dependent on well-tuned hyperparameters. This is particularly prominent in reinforcement learning (RL) where a small change in the configuration can lead to failure. Despite the importance of tuning hyperparameters, it remains expensive and is often done in a naive and laborious way. A recent solution to this problem is Population Based Training (PBT) which updates both weights and hyperparameters in a \emph{single training run} of a population of agents. PBT has been shown to be particularly effective in RL, leading to widespread use in the field. However, PBT lacks theoretical guarantees since it relies on random heuristics to explore the hyperparameter space. This inefficiency means it typically requires vast computational resources, which is prohibitive for many small and medium sized labs. In this work, we introduce the first provably efficient PBT-style algorithm, Population-Based Bandits (PB2). PB2 uses a probabilistic model to guide the search in an efficient way, making it possible to discover high performing hyperparameter configurations with far fewer agents than typically required by PBT. We show in a series of RL experiments that PB2 is able to achieve high performance with a modest computational budget.

[Understanding Global Feature Contributions With Additive Importance Measures](#)

- Ian Covert, Scott M. Lundberg, Su-In Lee
- abstract@[open-review](#): Understanding the inner workings of complex machine learning models is a long-standing problem and most recent research has focused on local interpretability. To assess the role of individual input features in a global sense, we explore the perspective of defining feature importance through the predictive power associated with each feature. We introduce two notions of predictive power (model-based and universal) and formalize this approach with a framework of additive importance measures, which unifies numerous methods in the literature. We then propose SAGE, a model-agnostic method that quantifies predictive power while accounting for feature interactions. Our experiments show that SAGE can be calculated efficiently and that it assigns more accurate importance values than other methods.

[Online Non-Convex Optimization with Imperfect Feedback](#)

- Amélie Héliou, Matthieu Martin, Panayotis Mertikopoulos, Thibaud Rahier
- abstract@[open-review](#): We consider the problem of online learning with non-convex losses. In terms of feedback, we assume that the learner observes “ or otherwise constructs “ an inexact model for the loss function encountered at each stage, and we propose a mixed-strategy learning policy based on dual averaging. In this general context, we derive a series of tight regret minimization guarantees, both for the learner’s static (external) regret, as well as the regret incurred against the best dynamic policy in hindsight. Subsequently, we apply this general template to the case where the learner only has access to the actual loss incurred at each stage of the process. This is achieved by means of a kernel-based estimator which generates an inexact model for each round’s loss function using only the learner’s realized losses as input.

[Co-Tuning for Transfer Learning](#)

- Kaichao You, Zhi Kou, Mingsheng Long, Jianmin Wang
- abstract@[open-review](#): Fine-tuning pre-trained deep neural networks (DNNs) to a target dataset, also known as transfer learning, is widely used in computer vision and NLP. Because task-specific layers mainly contain categorical information and categories vary with datasets, practitioners only partially transfer pre-trained models by discarding task-specific layers and fine-tuning bottom layers. However, it is a reckless loss to simply discard task-specific parameters who take up as many as \$20\%\$ of the total parameters in pre-trained models. To fully transfer pre-trained models, we propose a two-step framework named \textbf{Co-Tuning}: (i) learn the relationship between source categories and target categories from the pre-trained model and calibrated predictions; (ii) target labels (one-hot labels), as well as source labels (probabilistic labels) translated by the category relationship, collaboratively supervise the fine-tuning process. A simple instantiation of the framework shows strong empirical results in four visual classification tasks and one NLP classification task, bringing up to \$20\%\$ relative improvement. While state-of-the-art fine-tuning techniques mainly focus on how to impose regularization when data are not abundant, Co-Tuning works not only in medium-scale datasets (100 samples per class) but also in large-scale datasets (1000 samples per class) where regularization-based methods bring no gains over the vanilla fine-tuning. Co-Tuning relies on a typically valid assumption that the pre-trained dataset is diverse enough, implying its broad application area.

[Multifaceted Uncertainty Estimation for Label-Efficient Deep Learning](#)

- Weishi Shi, Xujiang Zhao, Feng Chen, Qi Yu
- abstract@[open-review](#): We present a novel multi-source uncertainty prediction approach that enables deep learning (DL) models to be actively trained with much less labeled data. By leveraging the second-order uncertainty representation provided by subjective logic (SL), we conduct evidence-based theoretical analysis and formally decompose the predicted entropy over multiple classes into two distinct sources of uncertainty: vacuity and dissonance, caused by lack of evidence and conflict of strong evidence, respectively. The evidence based entropy decomposition provides deeper insights on the nature of uncertainty, which can help effectively explore a large and high-dimensional unlabeled data space. We develop a novel loss function that augments DL based evidence prediction with uncertainty anchor sample identification. The accurately estimated multiple sources of uncertainty are systematically integrated and dynamically balanced using a data sampling function for label-efficient active deep learning (ADL). Experiments conducted over both synthetic and real data and comparison with competitive AL methods demonstrate the effectiveness of the proposed ADL model.

[Continuous Surface Embeddings](#)

- Natalia Neverova, David Novotny, Marc Szafraniec, Vasil Khalidov, Patrick Labatut, Andrea Vedaldi
- abstract@[open-review](#): In this work, we focus on the task of learning and representing dense correspondences in deformable object categories. While this problem has been considered before, solutions so far have been rather ad-hoc for specific object types (i.e., humans), often with significant manual work involved. However, scaling the geometry understanding to all objects in nature requires more automated approaches that can also express correspondences between related, but geometrically different objects. To this end, we propose a new, learnable image-based representation of dense correspondences. Our model predicts, for each pixel in a 2D image, an embedding vector of the corresponding vertex in the object mesh, therefore establishing dense correspondences between image pixels and 3D object geometry. We demonstrate that the proposed approach performs on par or better than the state-of-the-art methods for dense pose estimation for humans, while being conceptually simpler. We also collect a new in-the-wild dataset of dense correspondences for animal classes and demonstrate that our framework scales naturally to the new deformable object categories.

[Succinct and Robust Multi-Agent Communication With Temporal Message Control](#)

- Sai Qian Zhang, Qi Zhang, Jieyu Lin

- abstract@[open-review](#): Recent studies have shown that introducing communication between agents can significantly improve overall performance in cooperative Multi-agent reinforcement learning (MARL). However, existing communication schemes often require agents to exchange an excessive number of messages at run-time under a reliable communication channel, which hinders its practicality in many real-world situations. In this paper, we present \textit{Temporal Message Control} (TMC), a simple yet effective approach for achieving succinct and robust communication in MARL. TMC applies a temporal smoothing technique to drastically reduce the amount of information exchanged between agents. Experiments show that TMC can significantly reduce inter-agent communication overhead without impacting accuracy. Furthermore, TMC demonstrates much better robustness against transmission loss than existing approaches in lossy networking environments.

[Big Bird: Transformers for Longer Sequences](#)

- Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, Amr Ahmed
- abstract@[open-review](#): Transformers-based models, such as BERT, have been one of the most successful deep learning models for NLP. Unfortunately, one of their core limitations is the quadratic dependency (mainly in terms of memory) on the sequence length due to their full attention mechanism. To remedy this, we propose, BigBird, a sparse attention mechanism that reduces this quadratic dependency to linear. We show that BigBird is a universal approximator of sequence functions and is Turing complete, thereby preserving these properties of the quadratic, full attention model. Along the way, our theoretical analysis reveals some of the benefits of having $O(1)$ global tokens (such as CLS), that attend to the entire sequence as part of the sparse attention mechanism. The proposed sparse attention can handle sequences of length up to 8x of what was previously possible using similar hardware. As a consequence of the capability to handle longer context, BigBird drastically improves performance on various NLP tasks such as question answering and summarization. We also propose novel applications to genomics data.

[Neural Execution Engines: Learning to Execute Subroutines](#)

- Yujun Yan, Kevin Swersky, Danai Koutra, Parthasarathy Ranganathan, Milad Hashemi
- abstract@[open-review](#): A significant effort has been made to train neural networks that replicate algorithmic reasoning, but they often fail to learn the abstract concepts underlying these algorithms. This is evidenced by their inability to generalize to data distributions that are outside of their restricted training sets, namely larger inputs and unseen data. We study these generalization issues at the level of numerical subroutines that comprise common algorithms like sorting, shortest paths, and minimum spanning trees. First, we observe that transformer-based sequence-to-sequence models can learn subroutines like sorting a list of numbers, but their performance rapidly degrades as the length of lists grows beyond those found in the training set. We demonstrate that this is due to attention weights that lose fidelity with longer sequences, particularly when the input numbers are numerically similar. To address the issue, we propose a learned conditional masking mechanism, which enables the model to strongly generalize far outside of its training range with near-perfect accuracy on a variety of algorithms. Second, to generalize to unseen data, we show that encoding numbers with a binary representation leads to embeddings with rich structure once trained on downstream tasks like addition or multiplication. This allows the embedding to handle missing data by faithfully interpolating numbers not seen during training.

[Random Reshuffling: Simple Analysis with Vast Improvements](#)

- Konstantin Mishchenko, Ahmed Khaled, Peter Richtarik
- abstract@[open-review](#): Random Reshuffling (RR) is an algorithm for minimizing finite-sum functions that utilizes iterative gradient descent steps in conjunction with data reshuffling. Often contrasted with its sibling Stochastic Gradient Descent (SGD), RR is usually faster in practice and enjoys significant popularity in convex and non-convex optimization. The convergence rate of RR has attracted substantial attention recently and, for strongly convex and smooth functions, it was shown to converge faster than SGD if 1) the stepsize is small, 2) the gradients are bounded, and 3) the number of epochs is large. We remove these 3 assumptions, improve the dependence on the condition number from κ^2 to κ (resp. from κ to $\sqrt{\kappa}$) and, in addition, show that RR has a different type of variance. We argue through theory and experiments that the new variance type gives an additional justification of the superior performance of RR. To go beyond strong convexity, we present several results for non-strongly convex and non-convex objectives. We show that in all cases, our theory improves upon existing literature. Finally, we prove fast convergence of the Shuffle-Once (SO) algorithm, which shuffles the data only once, at the beginning of the optimization process. Our theory for strongly convex objectives tightly matches the known lower bounds for both RR and SO and substantiates the common practical heuristic of shuffling once or only a few times. As a byproduct of our analysis, we also get new results for the Incremental Gradient algorithm (IG), which does not shuffle the data at all.

[Long-Horizon Visual Planning with Goal-Conditioned Hierarchical Predictors](#)

- Karl Pertsch, Oleh Rybkin, Frederik Ebert, Shenghao Zhou, Dinesh Jayaraman, Chelsea Finn, Sergey Levine
- abstract@[open-review](#): The ability to predict and plan into the future is fundamental for agents acting in the world. To reach a faraway goal, we predict trajectories at multiple timescales, first devising a coarse plan towards the goal and then gradually filling in details. In contrast, current learning approaches for visual prediction and planning fail on long-horizon tasks as they generate predictions (1)~without considering goal information, and (2)~at the finest temporal resolution, one step at a time. In this work we propose a framework for visual prediction and planning that is able to overcome both of these limitations. First, we formulate the problem of predicting towards a goal and propose the corresponding class of latent space goal-conditioned predictors (GCPs). GCPs significantly improve planning efficiency by constraining the search space to only those trajectories that reach the goal. Further, we show how GCPs can be naturally formulated as hierarchical models that, given two observations, predict an observation between them, and by recursively subdividing each part of the trajectory generate complete sequences. This divide-and-conquer strategy is effective at long-term prediction, and enables us to design an effective hierarchical planning algorithm that optimizes trajectories in a coarse-to-fine manner. We show that by using both goal-conditioning and hierarchical prediction, GCPs enable us to solve visual planning tasks with much longer horizon than previously possible. See prediction and planning videos on the supplementary website: sites.google.com/view/video-gcp.

[Statistical Optimal Transport posed as Learning Kernel Embedding](#)

- Saketha Nath Jagarlapudi, Pratik Kumar Jawanpuria
- abstract@[open-review](#): The objective in statistical Optimal Transport (OT) is to consistently estimate the optimal transport plan/map solely using samples from the given source and target marginal distributions. This work takes the novel approach of posing statistical OT as that of learning the transport plan's kernel mean embedding from sample based estimates of marginal embeddings. The proposed estimator controls overfitting by employing maximum mean discrepancy based regularization, which is complementary to ϕ -divergence (entropy) based regularization popularly employed in existing estimators. A key result is that, under very mild conditions, ϵ -optimal recovery of the transport plan as well as the Barycentric-projection based transport map is possible with a sample complexity that is completely dimension-free. Moreover, the implicit smoothing in the kernel mean embeddings enables out-of-sample estimation. An appropriate representer theorem is proved leading to a kernelized convex formulation for the estimator, which can then be potentially used to perform OT even in non-standard domains. Empirical results illustrate the efficacy of the proposed approach.

[Dual-Resolution Correspondence Networks](#)

- Xinghui Li, Kai Han, Shuda Li, Victor Prisacariu
- abstract@[open-review](#): We tackle the problem of establishing dense pixel-wise correspondences between a pair of images. In this work, we introduce Dual-Resolution Correspondence Networks (DualRC-Net), to obtain pixel-wise correspondences in a coarse-to-fine manner. DualRC-Net extracts both

coarse- and fine- resolution feature maps. The coarse maps are used to produce a full but coarse 4D correlation tensor, which is then refined by a learnable neighbourhood consensus module. The fine-resolution feature maps are used to obtain the final dense correspondences guided by the refined coarse 4D correlation tensor. The selected coarse-resolution matching scores allow the fine-resolution features to focus only on a limited number of possible matches with high confidence. In this way, DualRC-Net dramatically increases matching reliability and localisation accuracy, while avoiding to apply the expensive 4D convolution kernels on fine-resolution feature maps. We comprehensively evaluate our method on large-scale public benchmarks including HPatches, InLoc, and Aachen Day-Night. It achieves state-of-the-art results on all of them.

[Advances in Black-Box VI: Normalizing Flows, Importance Weighting, and Optimization](#)

- Abhinav Agrawal, Daniel R. Sheldon, Justin Domke
- abstract@[open-review](#): Recent research has seen several advances relevant to black-box VI, but the current state of automatic posterior inference is unclear. One such advance is the use of normalizing flows to define flexible posterior densities for deep latent variable models. Another direction is the integration of Monte-Carlo methods to serve two purposes; first, to obtain tighter variational objectives for optimization, and second, to define enriched variational families through sampling. However, both flows and variational Monte-Carlo methods remain relatively unexplored for black-box VI. Moreover, on a pragmatic front, there are several optimization considerations like step-size scheme, parameter initialization, and choice of gradient estimators, for which there are no clear guidance in the existing literature. In this paper, we postulate that black-box VI is best addressed through a careful combination of numerous algorithmic components. We evaluate components relating to optimization, flows, and Monte-Carlo methods on a benchmark of 30 models from the Stan model library. The combination of these algorithmic components significantly advances the state-of-the-art "out of the box" variational inference.

[f-Divergence Variational Inference](#)

- Neng Wan, Dapeng Li, NAIRA HOVAKIMYAN
- abstract@[open-review](#): This paper introduces the f-divergence variational inference (f-VI) that generalizes variational inference to all f-divergences. Initiated from minimizing a crafty surrogate f-divergence that shares the statistical consistency with the f-divergence, the f-VI framework not only unifies a number of existing VI methods, e.g. Kullback–Leibler VI, Renyi's alpha-VI, and chi-VI, but offers a standardized toolkit for VI subject to arbitrary divergences from f-divergence family. A general f-variational bound is derived and provides a sandwich estimate of marginal likelihood (or evidence). The development of the f-VI unfolds with a stochastic optimization scheme that utilizes the reparameterization trick, importance weighting and Monte Carlo approximation; a mean-field approximation scheme that generalizes the well-known coordinate ascent variational inference (CAVI) is also proposed for f-VI. Empirical examples, including variational autoencoders and Bayesian neural networks, are provided to demonstrate the effectiveness and the wide applicability of f-VI.

[Unfolding recurrence by Greenâ€™s functions for optimized reservoir computing](#)

- Sandra Nestler, Christian Keup, David Dahmen, Matthieu Gilson, Holger Rauhut, Moritz Helias
- abstract@[open-review](#): Cortical networks are strongly recurrent, and neurons have intrinsic temporal dynamics. This sets them apart from deep feed-forward networks. Despite the tremendous progress in the application of deep feed-forward networks and their theoretical understanding, it remains unclear how the interplay of recurrence and non-linearities in recurrent cortical networks contributes to their function. The purpose of this work is to present a solvable recurrent network model that links to feed forward networks. By perturbative methods we transform the time-continuous, recurrent dynamics into an effective feed-forward structure of linear and non-linear temporal kernels. The resulting analytical expressions allow us to build optimal time-series classifiers from random reservoir networks. Firstly, this allows us to optimize not only the readout vectors, but also the input projection, demonstrating a strong potential performance gain. Secondly, the analysis exposes how the second order stimulus statistics is a crucial element that interacts with the non-linearity of the dynamics and boosts performance.

[The Dilemma of TriHard Loss and an Element-Weighted TriHard Loss for Person Re-Identification](#)

- Yihao Lv, Youzhi Gu, Liu Xinggao
- abstract@[open-review](#): Triplet loss with batch hard mining (TriHard loss) is an important variation of triplet loss inspired by the idea that hard triplets improve the performance of metric leaning networks. However, there is a dilemma in the training process. The hard negative samples contain various quite similar characteristics compared with anchors and positive samples in a batch. Features of these characteristics should be clustered between anchors and positive samples while are also utilized to repel between anchors and hard negative samples. It is harmful for learning mutual features within classes. Several methods to alleviate the dilemma are designed and tested. In the meanwhile, an element-weighted TriHard loss is emphatically proposed to enlarge the distance between partial elements of feature vectors selectively which represent the different characteristics between anchors and hard negative samples. Extensive evaluations are conducted on Market1501 and MSMT17 datasets and the results achieve state-of-the-art on public baselines.

[Disentangling by Subspace Diffusion](#)

- David Pfau, Irina Higgins, Alex Botev, SÃ©bastien RacaniÃ¨re
- abstract@[open-review](#): We present a novel nonparametric algorithm for symmetry-based disentangling of data manifolds, the Geometric Manifold Component Estimator (GEOMANCER). GEOMANCER provides a partial answer to the question posed by Higgins et al.(2018): is it possible to learn how to factorize a Lie group solely from observations of the orbit of an object it acts on? We show that fully unsupervised factorization of a data manifold is possible if the true metric of the manifold is known and each factor manifold has nontrivial holonomy – for example, rotation in 3D. Our algorithm works by estimating the subspaces that are invariant under random walk diffusion, giving an approximation to the de Rham decomposition from differential geometry. We demonstrate the efficacy of GEOMANCER on several complex synthetic manifolds. Our work reduces the question of whether unsupervised disentangling is possible to the question of whether unsupervised metric learning is possible, providing a unifying insight into the geometric nature of representation learning.

[Towards Neural Programming Interfaces](#)

- Zachary Brown, Nathaniel Robinson, David Wingate, Nancy Fulda
- abstract@[open-review](#): It is notoriously difficult to control the behavior of artificial neural networks such as generative neural language models. We recast the problem of controlling natural language generation as that of learning to interface with a pretrained language model, just as Application Programming Interfaces (APIs) control the behavior of programs by altering hyperparameters. In this new paradigm, a specialized neural network (called a Neural Programming Interface or NPI) learns to interface with a pretrained language model by manipulating the hidden activations of the pretrained model to produce desired outputs. Importantly, no permanent changes are made to the weights of the original model, allowing us to re-purpose pretrained models for new tasks without overwriting any aspect of the language model. We also contribute a new data set construction algorithm and GAN-inspired loss function that allows us to train NPI models to control outputs of autoregressive transformers. In experiments against other state-of-the-art approaches, we demonstrate the efficacy of our methods using OpenAIâ€™s GPT-2 model, successfully controlling noun selection, topic aversion, offensive speech filtering, and other aspects of language while largely maintaining the controlled model's fluency under deterministic settings.

[Discovering Symbolic Models from Deep Learning with Inductive Biases](#)

- Miles Cranmer, Alvaro Sanchez Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, Shirley Ho
- abstract@[open-review](#): We develop a general approach to distill symbolic representations of a learned deep model by introducing strong inductive biases. We focus on Graph Neural Networks (GNNs). The technique works as follows: we first encourage sparse latent representations when we train a GNN in a supervised setting, then we apply symbolic regression to components of the learned model to extract explicit physical relations. We find the correct known equations, including force laws and Hamiltonians, can be extracted from the neural network. We then apply our method to a non-trivial cosmology example—a detailed dark matter simulation—and discover a new analytic formula which can predict the concentration of dark matter from the mass distribution of nearby cosmic structures. The symbolic expressions extracted from the GNN using our technique also generalized to out-of-distribution-data better than the GNN itself. Our approach offers alternative directions for interpreting neural networks and discovering novel physical principles from the representations they learn.

[Real World Games Look Like Spinning Tops](#)

- Wojciech M. Czarnecki, Gauthier Gidel, Brendan Tracey, Karl Tuyls, Shayegan Omidshafiei, David Balduzzi, Max Jaderberg
- abstract@[open-review](#): This paper investigates the geometrical properties of real world games (e.g. Tic-Tac-Toe, Go, StarCraft II). We hypothesize that their geometrical structure resembles a spinning top, with the upright axis representing transitive strength, and the radial axis representing the non-transitive dimension, which corresponds to the number of cycles that exist at a particular transitive strength. We prove the existence of this geometry for a wide class of real world games by exposing their temporal nature. Additionally, we show that this unique structure also has consequences for learning - it clarifies why populations of strategies are necessary for training of agents, and how population size relates to the structure of the game. Finally, we empirically validate these claims by using a selection of nine real world two-player zero-sum symmetric games, showing 1) the spinning top structure is revealed and can be easily reconstructed by using a new method of Nash clustering to measure the interaction between transitive and cyclical strategy behaviour, and 2) the effect that population size has on the convergence of learning in these games.

[Cooperative Heterogeneous Deep Reinforcement Learning](#)

- Han Zheng, Pengfei Wei, Jing Jiang, Guodong Long, Qinghua Lu, Chengqi Zhang
- abstract@[open-review](#): Numerous deep reinforcement learning agents have been proposed, and each of them has its strengths and flaws. In this work, we present a Cooperative Heterogeneous Deep Reinforcement Learning (CHDRL) framework that can learn a policy by integrating the advantages of heterogeneous agents. Specifically, we propose a cooperative learning framework that classifies heterogeneous agents into two classes: global agents and local agents. Global agents are off-policy agents that can utilize experiences from the other agents. Local agents are either on-policy agents or population-based evolutionary algorithms (EAs) agents that can explore the local area effectively. We employ global agents, which are sample-efficient, to guide the learning of local agents so that local agents can benefit from the sample-efficient agents and simultaneously maintain their advantages, e.g., stability. Global agents also benefit from effective local searches. Experimental studies on a range of continuous control tasks from the Mujoco benchmark show that CHDRL achieves better performance compared with state-of-the-art baselines.

[Mitigating Forgetting in Online Continual Learning via Instance-Aware Parameterization](#)

- Hung-Jen Chen, An-Chieh Cheng, Da-Cheng Juan, Wei Wei, Min Sun
- abstract@[open-review](#): Online continual learning is a challenging scenario where a model needs to learn from a continuous stream of data without revisiting any previously encountered data instances. The phenomenon of catastrophic forgetting is worsened since the model should not only address the forgetting at the task-level but also at the data instance-level within the same task. To mitigate this, we leverage the concept of "instance awareness" in the neural network, where each data instance is classified by a path in the network searched by the controller from a meta-graph. To preserve the knowledge we learn from previous instances, we proposed a method to protect the path by restricting the gradient updates of one instance from overriding past updates calculated from previous instances if these instances are not similar. On the other hand, it also encourages fine-tuning the path if the incoming instance shares the similarity with previous instances. The mechanism of selecting paths according to instances similarity is naturally determined by the controller, which is compact and online updated. Experimental results show that the proposed method outperforms state-of-the-arts in online continual learning. Furthermore, the proposed method is evaluated against a realistic setting where the boundaries between tasks are blurred. Experimental results confirm that the proposed method outperforms the state-of-the-arts on CIFAR-10, CIFAR-100, and Tiny-ImageNet.

[ImpatientCapsAndRuns: Approximately Optimal Algorithm Configuration from an Infinite Pool](#)

- Gellert Weisz, András György, Wei-I Lin, Devon Graham, Kevin Leyton-Brown, Csaba Szepesvári, Brendan Lucier
- abstract@[open-review](#): Algorithm configuration procedures optimize parameters of a given algorithm to perform well over a distribution of inputs. Recent theoretical work focused on the case of selecting between a small number of alternatives. In practice, parameter spaces are often very large or infinite, and so successful heuristic procedures discard parameters “impatiently”, based on very few observations. Inspired by this idea, we introduce ImpatientCapsAndRuns, which quickly discards less promising configurations, significantly speeding up the search procedure compared to previous algorithms with theoretical guarantees, while still achieving optimal runtime up to logarithmic factors under mild assumptions. Experimental results demonstrate a practical improvement.

[Dense Correspondences between Human Bodies via Learning Transformation Synchronization on Graphs](#)

- Xiangru Huang, Haitao Yang, Etienne Vouga, Qixing Huang
- abstract@[open-review](#): We introduce an approach for establishing dense correspondences between partial scans of human models and a complete template model. Our approach's key novelty lies in formulating dense correspondence computation as initializing and synchronizing local transformations between the scan and the template model. We introduce an optimization formulation for synchronizing transformations among a graph of the input scan, which automatically enforces smoothness of correspondences and recovers the underlying articulated deformations. We then show how to convert the iterative optimization procedure among a graph of the input scan into an end-to-end trainable network. The network design utilizes additional trainable parameters to break the barrier of the original optimization formulation's exact and robust recovery conditions. Experimental results on benchmark datasets demonstrate that our approach considerably outperforms baseline approaches in accuracy and robustness.

[Reasoning about Uncertainties in Discrete-Time Dynamical Systems using Polynomial Forms.](#)

- Sriram Sankaranarayanan, Yi Chou, Eric Goubault, Sylvie Putot
- abstract@[open-review](#): In this paper, we propose polynomial forms to represent distributions of state variables over time for discrete-time stochastic dynamical systems. This problem arises in a variety of applications in areas ranging from biology to robotics. Our approach allows us to rigorously represent the probability distribution of state variables over time, and provide guaranteed bounds on the expectations, moments and probabilities of tail events involving the state variables. First, we recall ideas from interval arithmetic, and use them to rigorously represent the state variables at time t as a function of the initial state variables and noise symbols that model the random exogenous inputs encountered before time t . Next, we show how concentration of measure inequalities can be employed to prove rigorous bounds on the tail probabilities of these state variables. We demonstrate interesting applications that demonstrate how our approach can be useful in some situations to establish mathematically guaranteed bounds that are of a different nature from those obtained through simulations with pseudo-random numbers.

[Applications of Common Entropy for Causal Inference](#)

- Murat Kocaoglu, Sanjay Shakkottai, Alexandros G. Dimakis, Constantine Caramanis, Sriram Vishwanath
- abstract@[open-review](#): We study the problem of discovering the simplest latent variable that can make two observed discrete variables conditionally independent. The minimum entropy required for such a latent is known as common entropy in information theory. We extend this notion to Renyi common entropy by minimizing the Renyi entropy of the latent variable. To efficiently compute common entropy, we propose an iterative algorithm that can be used to discover the trade-off between the entropy of the latent variable and the conditional mutual information of the observed variables. We show two applications of common entropy in causal inference: First, under the assumption that there are no low-entropy mediators, it can be used to distinguish direct causation from spurious correlation among almost all joint distributions on simple causal graphs with two observed variables. Second, common entropy can be used to improve constraint-based methods such as PC or FCI algorithms in the small-sample regime, where these methods are known to struggle. We propose a modification to these constraint-based methods to assess if a separating set found by these algorithms are valid using common entropy. We finally evaluate our algorithms on synthetic and real data to establish their performance.

[SGD with shuffling: optimal rates without component convexity and large epoch requirements](#)

- Kwangjun Ahn, Chulhee Yun, Suvrit Sra
- abstract@[open-review](#): We study without-replacement SGD for solving finite-sum optimization problems. Specifically, depending on how the indices of the finite-sum are shuffled, we consider the RandomShuffle (shuffle at the beginning of each epoch) and SingleShuffle (shuffle only once) algorithms. First, we establish minimax optimal convergence rates of these algorithms up to poly-log factors. Notably, our analysis is general enough to cover gradient dominated nonconvex costs, and does not rely on the convexity of individual component functions unlike existing optimal convergence results. Secondly, assuming convexity of the individual components, we further sharpen the tight convergence results for RandomShuffle by removing the drawbacks common to all prior arts: large number of epochs required for the results to hold, and extra poly-log factor gaps to the lower bound.

[Unsupervised Joint k-node Graph Representations with Compositional Energy-Based Models](#)

- Leonardo Cotta, Carlos H. C. Teixeira, Ananthram Swami, Bruno Ribeiro
- abstract@[open-review](#): Existing Graph Neural Network (GNN) methods that learn inductive unsupervised graph representations focus on learning node and edge representations by predicting observed edges in the graph. Although such approaches have shown advances in downstream node classification tasks, they are ineffective in jointly representing larger k-node sets, $k \geq 2$. We propose MHM-GNN, an inductive unsupervised graph representation approach that combines joint k-node representations with energy-based models (hypergraph Markov networks) and GNNs. To address the intractability of the loss that arises from this combination, we endow our optimization with a loss upper bound using a finite-sample unbiased Markov Chain Monte Carlo estimator. Our experiments show that the unsupervised joint k-node representations of MHM-GNN produce better unsupervised representations than existing approaches from the literature.

[Neural Manifold Ordinary Differential Equations](#)

- Aaron Lou, Derek Lim, Isay Katsman, Leo Huang, Qingxuan Jiang, Ser Nam Lim, Christopher M. De Sa
- abstract@[open-review](#): To better conform to data geometry, recent deep generative modelling techniques adapt Euclidean constructions to non-Euclidean spaces. In this paper, we study normalizing flows on manifolds. Previous work has developed flow models for specific cases; however, these advancements hand craft layers on a manifold-by-manifold basis, restricting generality and inducing cumbersome design constraints. We overcome these issues by introducing Neural Manifold Ordinary Differential Equations, a manifold generalization of Neural ODEs, which enables the construction of Manifold Continuous Normalizing Flows (MCNFs). MCNFs require only local geometry (therefore generalizing to arbitrary manifolds) and compute probabilities with continuous change of variables (allowing for a simple and expressive flow construction). We find that leveraging continuous manifold dynamics produces a marked improvement for both density estimation and downstream tasks.

[CO-Optimal Transport](#)

- Vayer Titouan, Ievgen Redko, RÃ©mi Flamary, Nicolas Courty
- abstract@[open-review](#): Optimal transport (OT) is a powerful geometric and probabilistic tool for finding correspondences and measuring similarity between two distributions. Yet, its original formulation relies on the existence of a cost function between the samples of the two distributions, which makes it impractical when they are supported on different spaces. To circumvent this limitation, we propose a novel OT problem, named COOT for CO-Optimal Transport, that simultaneously optimizes two transport maps between both samples and features, contrary to other approaches that either discard the individual features by focusing on pairwise distances between samples or need to model explicitly the relations between them. We provide a thorough theoretical analysis of our problem, establish its rich connections with other OT-based distances and demonstrate its versatility with two machine learning applications in heterogeneous domain adaptation and co-clustering/data summarization, where COOT leads to performance improvements over the state-of-the-art methods.

[Continuous Meta-Learning without Tasks](#)

- James Harrison, Apoorva Sharma, Chelsea Finn, Marco Pavone
- abstract@[open-review](#): Meta-learning is a promising strategy for learning to efficiently learn using data gathered from a distribution of tasks. However, the meta-learning literature thus far has focused on the task segmented setting, where at train-time, offline data is assumed to be split according to the underlying task, and at test-time, the algorithms are optimized to learn in a single task. In this work, we enable the application of generic meta-learning algorithms to settings where this task segmentation is unavailable, such as continual online learning with unsegmented time series data. We present meta-learning via online changepoint analysis (MOCA), an approach which augments a meta-learning algorithm with a differentiable Bayesian changepoint detection scheme. The framework allows both training and testing directly on time series data without segmenting it into discrete tasks. We demonstrate the utility of this approach on three nonlinear meta-regression benchmarks as well as two meta-image-classification benchmarks.

[A mathematical theory of cooperative communication](#)

- Pei Wang, Junqi Wang, Pushpi Parannama, Patrick Shafto
- abstract@[open-review](#): Cooperative communication plays a central role in theories of human cognition, language, development, culture, and human-robot interaction. Prior models of cooperative communication are algorithmic in nature and do not shed light on why cooperation may yield effective belief transmission and what limitations may arise due to differences between beliefs of agents. Through a connection to the theory of optimal transport, we establish a mathematical framework for cooperative communication. We derive prior models as special cases, statistical interpretations of belief transfer plans, and proofs of robustness and instability. Computational simulations support and elaborate our theoretical results, and demonstrate fit to human behavior. The results show that cooperative communication provably enables effective, robust belief transmission which is required to explain feats of human learning and improve human-machine interaction.

[Penalized Langevin dynamics with vanishing penalty for smooth and log-concave targets](#)

- Avetik Karagulyan, Arnak Dalalyan
- abstract@[open-review](#): We study the problem of sampling from a probability distribution on \mathbb{R}^p defined via a convex and smooth potential function.

We first consider a continuous-time diffusion-type process, termed Penalized Langevin dynamics (PLD), the drift of which is the negative gradient of the potential plus a linear penalty that vanishes when time goes to infinity. An upper bound on the Wasserstein-2 distance between the distribution of the PLD at time t and the target is established.

This upper bound highlights the influence of the speed of decay of the penalty on the accuracy of approximation. As a consequence, in the case of low-temperature limit we infer a new result on the convergence of the penalized gradient flow for the optimization problem.

[Learning Invariances in Neural Networks from Training Data](#)

- Gregory Benton, Marc Finzi, Pavel Izmailov, Andrew G. Wilson
- abstract@[open-review](#): Invariances to translations have imbued convolutional neural networks with powerful generalization properties. However, we often do not know a priori what invariances are present in the data, or to what extent a model should be invariant to a given augmentation. We show how to learn invariances by parameterizing a distribution over augmentations and optimizing the training loss simultaneously with respect to the network parameters and augmentation parameters. With this simple procedure we can recover the correct set and extent of invariances on image classification, regression, segmentation, and molecular property prediction from a large space of augmentations, on training data alone. We show our approach is competitive with methods that are specialized to each task with the appropriate hard-coded invariances, without providing any prior knowledge of which invariance is needed.

[A Finite-Time Analysis of Two Time-Scale Actor-Critic Methods](#)

- Yue Frank Wu, Weitong ZHANG, Pan Xu, Quanquan Gu
- abstract@[open-review](#): Actor-critic (AC) methods have exhibited great empirical success compared with other reinforcement learning algorithms, where the actor uses the policy gradient to improve the learning policy and the critic uses temporal difference learning to estimate the policy gradient. Under the two time-scale learning rate schedule, the asymptotic convergence of AC has been well studied in the literature. However, the non-asymptotic convergence and finite sample complexity of actor-critic methods are largely open. In this work, we provide a non-asymptotic analysis for two time-scale actor-critic methods under non-i.i.d. setting. We prove that the actor-critic method is guaranteed to find a first-order stationary point (i.e., $\|\nabla J(\theta)\|_2^2 \leq \epsilon$) of the non-concave performance function $J(\theta)$, with $\mathcal{O}(\epsilon^{-2.5})$ sample complexity. To the best of our knowledge, this is the first work providing finite-time analysis and sample complexity bound for two time-scale actor-critic methods.

[Pruning Filter in Filter](#)

- Fanxu Meng, Hao Cheng, Ke Li, Huixiang Luo, Xiaowei Guo, Guangming Lu, Xing Sun
- abstract@[open-review](#): Pruning has become a very powerful and effective technique to compress and accelerate modern neural networks. Existing pruning methods can be grouped into two categories: filter pruning (FP) and weight pruning (WP). FP wins at hardware compatibility but loses at the compression ratio compared with WP. To converge the strength of both methods, we propose to prune the filter in the filter. Specifically, we treat a filter F, whose size is CKK, as KK stripes, i.e., 11 filters, then by pruning the stripes instead of the whole filter, we can achieve finer granularity than traditional FP while being hardware friendly. We term our method as SWP (Stripe-Wise Pruning). SWP is implemented by introducing a novel learnable matrix called Filter Skeleton, whose values reflect the optimal shape of each filter. As some recent work has shown that the pruned architecture is more crucial than the inherited important weights, we argue that the architecture of a single filter, i.e., the Filter Skeleton, also matters. Through extensive experiments, we demonstrate that SWP is more effective compared to the previous FP-based methods and achieves the state-of-art pruning ratio on CIFAR-10 and ImageNet datasets without obvious accuracy drop.

[Learning to Mutate with Hypergradient Guided Population](#)

- Zhiqiang Tao, Yaliang Li, Bolin Ding, Ce Zhang, Jingren Zhou, Yun Fu
- abstract@[open-review](#): Computing the gradient of model hyperparameters, i.e., hypergradient, enables a promising and natural way to solve the hyperparameter optimization task. However, gradient-based methods could lead to suboptimal solutions due to the non-convex nature of optimization in a complex hyperparameter space. In this study, we propose a hyperparameter mutation (HPM) algorithm to explicitly consider a learnable trade-off between using global and local search, where we adopt a population of student models to simultaneously explore the hyperparameter space guided by hypergradient and leverage a teacher model to mutate the underperforming students by exploiting the top ones. The teacher model is implemented with an attention mechanism and is used to learn a mutation schedule for different hyperparameters on the fly. Empirical evidence on synthetic functions is provided to show that HPM outperforms hypergradient significantly. Experiments on two benchmark datasets are also conducted to validate the effectiveness of the proposed HPM algorithm for training deep neural networks compared with several strong baselines.

[A convex optimization formulation for multivariate regression](#)

- Yunzhang Zhu
- abstract@[open-review](#): Multivariate regression (or multi-task learning) concerns the task of predicting the value of multiple responses from a set of covariates. In this article, we propose a convex optimization formulation for high-dimensional multivariate linear regression under a general error covariance structure. The main difficulty with simultaneous estimation of the regression coefficients and the error covariance matrix lies in the fact that the negative log-likelihood function is not convex. To overcome this difficulty, a new parameterization is proposed, under which the negative log-likelihood function is proved to be convex. For faster computation, two other alternative loss functions are also considered, and proved to be convex under the proposed parameterization. This new parameterization is also useful for covariate-adjusted Gaussian graphical modeling in which the inverse of the error covariance matrix is of interest. A joint non-asymptotic analysis of the regression coefficients and the error covariance matrix is carried out under the new parameterization. In particular, we show that the proposed method recovers the oracle estimator under sharp scaling conditions, and rates of convergence in terms of vector $\|\cdot\|_\infty$ norm are also established. Empirically, the proposed methods outperform existing high-dimensional multivariate linear regression methods that are based on either minimizing certain non-convex criteria or certain two-step procedures.

[Online Meta-Critic Learning for Off-Policy Actor-Critic Methods](#)

- Wei Zhou, Yiying Li, Yongxin Yang, Huaimin Wang, Timothy Hospedales
- abstract@[open-review](#): Off-Policy Actor-Critic (OffP-AC) methods have proven successful in a variety of continuous control tasks. Normally, the critic's action-value function is updated using temporal-difference, and the critic in turn provides a loss for the actor that trains it to take actions with higher expected return. In this paper, we introduce a flexible and augmented meta-critic that observes the learning process and meta-learns an additional loss for the actor that accelerates and improves actor-critic learning. Compared to existing meta-learning algorithms, meta-critic is rapidly learned online for a single task, rather than slowly over a family of tasks. Crucially, our meta-critic is designed for off-policy based learners, which currently provide state-of-the-art reinforcement learning sample efficiency. We demonstrate that online meta-critic learning benefits to a variety of continuous control tasks when combined with contemporary OffP-AC methods DDPG, TD3 and SAC.

[The All-or-Nothing Phenomenon in Sparse Tensor PCA](#)

- Jonathan Niles-Weed, Ilias Zadik
- abstract@[open-review](#): We study the statistical problem of estimating a rank-one sparse tensor corrupted by additive gaussian noise, a Gaussian additive model also known as sparse tensor PCA. We show that for Bernoulli and Bernoulli-Rademacher distributed signals and \emph{for all} sparsity levels which are sublinear in the dimension of the signal, the sparse tensor PCA model exhibits a phase transition called the \emph{all-or-nothing phenomenon}. This is the property that for some signal-to-noise ratio (SNR) SNR_c and any fixed $\epsilon > 0$, if the SNR of the model is below $\left(1 - \epsilon\right)\mathrm{SNR}_c$, then it is impossible to achieve any arbitrarily small constant correlation with the hidden signal, while if the SNR is above $\left(1 + \epsilon\right)\mathrm{SNR}_c$, then it is possible to achieve almost perfect correlation with the hidden signal. The all-or-nothing phenomenon was initially established in the context of sparse linear regression, and over the last year also in the context of sparse 2-tensor (matrix) PCA and Bernoulli group testing. Our results follow from a more general result showing that for any Gaussian additive model with a discrete uniform prior, the all-or-nothing phenomenon follows as a direct outcome of an appropriately defined ``near-orthogonality'' property of the support of the prior distribution.

[Synthesize, Execute and Debug: Learning to Repair for Neural Program Synthesis](#)

- Kavi Gupta, Peter Ebert Christensen, Xinyun Chen, Dawn Song
- abstract@[open-review](#): The use of deep learning techniques has achieved significant progress for program synthesis from input-output examples. However, when the program semantics become more complex, it still remains a challenge to synthesize programs that are consistent with the specification. In this work, we propose SED, a neural program generation framework that incorporates synthesis, execution, and debugging stages. Instead of purely relying on the neural program synthesizer to generate the final program, SED first produces initial programs using the neural program synthesizer component, then utilizes a neural program debugger to iteratively repair the generated programs. The integration of the debugger component enables SED to modify the programs based on the execution results and specification, which resembles the coding process of human programmers. On Karel, a challenging input-output program synthesis benchmark, SED reduces the error rate of the neural program synthesizer itself by a considerable margin, and outperforms the standard beam search for decoding.

[ARMA Nets: Expanding Receptive Field for Dense Prediction](#)

- Jiahao Su, Shiqi Wang, Furong Huang
- abstract@[open-review](#): Global information is essential for dense prediction problems, whose goal is to compute a discrete or continuous label for each pixel in the images. Traditional convolutional layers in neural networks, initially designed for image classification, are restrictive in these problems since the filter size limits their receptive fields. In this work, we propose to replace any traditional convolutional layer with an autoregressive moving-average (ARMA) layer, a novel module with an adjustable receptive field controlled by the learnable autoregressive coefficients. Compared with traditional convolutional layers, our ARMA layer enables explicit interconnections of the output neurons and learns its receptive field by adapting the autoregressive coefficients of the interconnections. ARMA layer is adjustable to different types of tasks: for tasks where global information is crucial, it is capable of learning relatively large autoregressive coefficients to allow for an output neuron's receptive field covering the entire input; for tasks where only local information is required, it can learn small or near zero autoregressive coefficients and automatically reduces to a traditional convolutional layer. We show both theoretically and empirically that the effective receptive field of networks with ARMA layers (named ARMA networks) expands with larger autoregressive coefficients. We also provably solve the instability problem of learning and prediction in the ARMA layer through a re-parameterization mechanism. Additionally, we demonstrate that ARMA networks substantially improve their baselines on challenging dense prediction tasks, including video prediction and semantic segmentation.

[Diversity-Guided Multi-Objective Bayesian Optimization With Batch Evaluations](#)

- Mina Konakovic Lukovic, Yunsheng Tian, Wojciech Matusik
- abstract@[open-review](#): Many science, engineering, and design optimization problems require balancing the trade-offs between several conflicting objectives. The objectives are often black-box functions whose evaluations are time-consuming and costly. Multi-objective Bayesian optimization can be used to automate the process of discovering the set of optimal solutions, called Pareto-optimal, while minimizing the number of performed evaluations. To further reduce the evaluation time in the optimization process, testing of several samples in parallel can be deployed. We propose a novel multi-objective Bayesian optimization algorithm that iteratively selects the best batch of samples to be evaluated in parallel. Our algorithm approximates and analyzes a piecewise-continuous Pareto set representation. This representation allows us to introduce a batch selection strategy that optimizes for both hypervolume improvement and diversity of selected samples in order to efficiently advance promising regions of the Pareto front. Experiments on both synthetic test functions and real-world benchmark problems show that our algorithm predominantly outperforms relevant state-of-the-art methods. Code is available at <https://github.com/yunshengtian/DGEMO>.

[SOLOv2: Dynamic and Fast Instance Segmentation](#)

- Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, Chunhua Shen
- abstract@[open-review](#): In this work, we design a simple, direct, and fast framework for instance segmentation with strong performance. To this end, we propose a novel and effective approach, termed SOLOv2, following the principle of the SOLO method [32]. First, our new framework is empowered by an efficient and holistic instance mask representation scheme, which dynamically segments each instance in the image, without resorting to bounding box detection. Specifically, the object mask generation is decoupled into a mask kernel prediction and mask feature learning, which are responsible for generating convolution kernels and the feature maps to be convolved with, respectively. Second, SOLOv2 significantly reduces inference overhead with our novel matrix non-maximum suppression (NMS) technique. Our Matrix NMS performs NMS with parallel matrix operations in one shot, and yields better results. We demonstrate that the proposed SOLOv2 achieves the state-of-the-art performance with high efficiency, making it suitable for both mobile and cloud applications. A light-weight version of SOLOv2 executes at 31.3 FPS and yields 37.1% AP on COCO test-dev. Moreover, our state-of-the-art results in object detection (from our mask byproduct) and panoptic segmentation show the potential of SOLOv2 to serve as a new strong baseline for many instance-level recognition tasks. Code is available at <https://git.io/AdelaiDet>

[Robust Recovery via Implicit Bias of Discrepant Learning Rates for Double Over-parameterization](#)

- Chong You, Zhihui Zhu, Qing Qu, Yi Ma
- abstract@[open-review](#): Recent advances have shown that implicit bias of gradient descent on over-parameterized models enables the recovery of low-rank matrices from linear measurements, even with no prior knowledge on the intrinsic rank. In contrast, for \em robust low-rank matrix recovery from \em grossly corrupted measurements, over-parameterization leads to overfitting without prior knowledge on both the intrinsic rank and sparsity of corruption. This paper shows that with a \em double over-parameterization for both the low-rank matrix and sparse corruption, gradient descent with \em discrepant learning rates provably recovers the underlying matrix even without prior knowledge on neither rank of the matrix nor sparsity of the corruption. We further extend our approach for the robust recovery of natural images by over-parameterizing images with deep convolutional networks. Experiments show that our method handles different test images and varying corruption levels with a single learning pipeline where the network width and termination conditions do not need to be adjusted on a case-by-case basis. Underlying the success is again the implicit bias with discrepant learning rates on different over-parameterized parameters, which may bear on broader applications.

[Axioms for Learning from Pairwise Comparisons](#)

- Ritesh Noothigattu, Dominik Peters, Ariel D. Procaccia
- abstract@[open-review](#): To be well-behaved, systems that process preference data must satisfy certain conditions identified by economic decision theory and by social choice theory. In ML, preferences and rankings are commonly learned by fitting a probabilistic model to noisy preference data. The behavior of this learning process from the view of economic theory has previously been studied for the case where the data consists of rankings. In practice, it is more common to have only pairwise comparison data, and the formal properties of the associated learning problem are more challenging to analyze. We show that a large class of random utility models (including the Thurstoneâ€“Mosteller Model), when estimated using the MLE, satisfy a Pareto efficiency condition. These models also satisfy a strong monotonicity property, which implies that the learning process is responsive to input data. On the other hand, we show that these models fail certain other consistency conditions from social choice theory, and in particular do not always follow the majority opinion. Our results inform existing and future applications of random utility models for societal decision making.

[Continuous Regularized Wasserstein Barycenters](#)

- Lingxiao Li, Aude Genevay, Mikhail Yurochkin, Justin M. Solomon
- abstract@[open-review](#): Wasserstein barycenters provide a geometrically meaningful way to aggregate probability distributions, built on the theory of optimal transport. They are difficult to compute in practice, however, leading previous work to restrict their supports to finite sets of points. Leveraging a new dual formulation for the regularized Wasserstein barycenter problem, we introduce a stochastic algorithm that constructs a continuous approximation of the barycenter. We establish strong duality and use the corresponding primal-dual relationship to parametrize the barycenter implicitly using the dual potentials of regularized transport problems. The resulting problem can be solved with stochastic gradient descent, which yields an efficient online algorithm to approximate the barycenter of continuous distributions given sample access. We demonstrate the effectiveness of our approach and compare against previous work on synthetic examples and real-world applications.

[Spectral Temporal Graph Neural Network for Multivariate Time-series Forecasting](#)

- Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Congrui Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, Qi Zhang
- abstract@[open-review](#): Multivariate time-series forecasting plays a crucial role in many real-world applications. It is a challenging problem as one needs to consider both intra-series temporal correlations and inter-series correlations simultaneously. Recently, there have been multiple works trying to capture both correlations, but most, if not all of them only capture temporal correlations in the time domain and resort to pre-defined priors as inter-series relationships. In this paper, we propose Spectral Temporal Graph Neural Network (StemGNN) to further improve the accuracy of multivariate time-series forecasting. StemGNN captures inter-series correlations and temporal dependencies jointly in the spectral domain. It combines Graph Fourier Transform (GFT) which models inter-series correlations and Discrete Fourier Transform (DFT) which models temporal dependencies in an end-to-end framework. After passing through GFT and DFT, the spectral representations hold clear patterns and can be predicted effectively by convolution and sequential learning modules. Moreover, StemGNN learns inter-series correlations automatically from the data without using pre-defined priors. We conduct extensive experiments on ten real-world datasets to demonstrate the effectiveness of StemGNN.

[Online Multitask Learning with Long-Term Memory](#)

- Mark Herbster, Stephen Pasteris, Lisa Tse
- abstract@[open-review](#): We introduce a novel online multitask setting. In this setting each task is partitioned into a sequence of segments that is unknown to the learner. Associated with each segment is a hypothesis from some hypothesis class. We give algorithms that are designed to exploit the scenario where there are many such segments but significantly fewer associated hypotheses. We prove regret bounds that hold for any segmentation of the tasks and any association of hypotheses to the segments. In the single-task setting this is equivalent to switching with long-term memory in the sense of [Bousquet and Warmuth 2011]. We provide an algorithm that predicts on each trial in time linear in the number of hypotheses when the hypothesis class is finite. We also consider infinite hypothesis classes from reproducing kernel Hilbert spaces for which we give an algorithm whose per trial time complexity is cubic in the number of cumulative trials. In the single-task special case this is the first example of an efficient regret-bounded switching algorithm with long-term memory for a non-parametric hypothesis class.

[Fewer is More: A Deep Graph Metric Learning Perspective Using Fewer Proxies](#)

- Yuehua Zhu, Muli Yang, Cheng Deng, Wei Liu
- abstract@[open-review](#): Deep metric learning plays a key role in various machine learning tasks. Most of the previous works have been confined to sampling from a mini-batch, which cannot precisely characterize the global geometry of the embedding space. Although researchers have developed proxy- and classification-based methods to tackle the sampling issue, those methods inevitably incur a redundant computational cost. In this paper, we propose a novel Proxy-based deep Graph Metric Learning (ProxyGML) approach from the perspective of graph classification, which uses fewer proxies yet achieves better comprehensive performance. Specifically, multiple global proxies are leveraged to collectively approximate the original data points for each class. To efficiently capture local neighbor relationships, a small number of such proxies are adaptively selected to construct similarity subgraphs between these proxies and each data point. Further, we design a novel reverse label propagation algorithm, by which the neighbor relationships are adjusted according to ground-truth labels, so that a discriminative metric space can be learned during the process of subgraph classification. Extensive experiments carried out on widely-used CUB-200-2011, Cars196, and Stanford Online Products datasets demonstrate the superiority of the proposed ProxyGML over the state-of-the-art methods in terms of both effectiveness and efficiency. The source code is publicly available at \url{https://github.com/YuehuaZhu/ProxyGML}.

[Adaptive Graph Convolutional Recurrent Network for Traffic Forecasting](#)

- LEI BAI, Lina Yao, Can Li, Xianzhi Wang, Can Wang
- abstract@[open-review](#): Modeling complex spatial and temporal correlations in the correlated time series data is indispensable for understanding the traffic dynamics and predicting the future status of an evolving traffic system. Recent works focus on designing complicated graph neural network architectures to capture shared patterns with the help of pre-defined graphs. In this paper, we argue that learning node-specific patterns is essential for traffic forecasting while pre-defined graph is avoidable. To this end, we propose two adaptive modules for enhancing Graph Convolutional Network (GCN) with new capabilities: 1) a Node Adaptive Parameter Learning (NAPL) module to capture node-specific patterns; 2) a Data Adaptive Graph Generation (DAGG) module to infer the inter-dependencies among different traffic series automatically. We further propose an Adaptive Graph Convolutional Recurrent Network (AGCRN) to capture fine-grained spatial and temporal correlations in traffic series automatically based on the two modules and recurrent networks. Our experiments on two real-world traffic datasets show AGCRN outperforms state-of-the-art by a significant margin without pre-defined graphs about spatial connections.

[On Reward-Free Reinforcement Learning with Linear Function Approximation](#)

- Ruosong Wang, Simon S. Du, Lin Yang, Russ R. Salakhutdinov
- abstract@[open-review](#): Reward-free reinforcement learning (RL) is a framework which is suitable for both the batch RL setting and the setting where there are many reward functions of interest. During the exploration phase, an agent collects samples without using a pre-specified reward function. After the exploration phase, a reward function is given, and the agent uses samples collected during the exploration phase to compute a near-optimal policy. Jin et al. [2020] showed that in the tabular setting, the agent only needs to collect polynomial number of samples (in terms of the number states, the number of actions, and the planning horizon) for reward-free RL. However, in practice, the number of states and actions can be large, and thus function

approximation schemes are required for generalization. In this work, we give both positive and negative results for reward-free RL with linear function approximation. We give an algorithm for reward-free RL in the linear Markov decision process setting where both the transition and the reward admit linear representations. The sample complexity of our algorithm is polynomial in the feature dimension and the planning horizon, and is completely independent of the number of states and actions. We further give an exponential lower bound for reward-free RL in the setting where only the optimal Q\$-function admits a linear representation. Our results imply several interesting exponential separations on the sample complexity of reward-free RL.

[Robustness of Community Detection to Random Geometric Perturbations](#)

- Sandrine Peche, Vianney Perchet
- abstract@[open-review](#): We consider the stochastic block model where connection between vertices is perturbed by some latent (and unobserved) random geometric graph. The objective is to prove that spectral methods are robust to this type of noise, even if they are agnostic to the presence (or not) of the random graph. We provide explicit regimes where the second eigenvector of the adjacency matrix is highly correlated to the true community vector (and therefore when weak/exact recovery is possible). This is possible thanks to a detailed analysis of the spectrum of the latent random graph, of its own interest.

[Learning outside the Black-Box: The pursuit of interpretable models](#)

- Jonathan Crabbe, Yao Zhang, William Zame, Mihaela van der Schaar
- abstract@[open-review](#): Machine learning has proved its ability to produce accurate models -- but the deployment of these models outside the machine learning community has been hindered by the difficulties of interpreting these models. This paper proposes an algorithm that produces a continuous global interpretation of any given continuous black-box function. Our algorithm employs a variation of projection pursuit in which the ridge functions are chosen to be Meijer G-functions, rather than the usual polynomial splines. Because Meijer G-functions are differentiable in their parameters, we can "tune" the parameters of the representation by gradient descent; as a consequence, our algorithm is efficient. Using five familiar data sets from the UCI repository and two familiar machine learning algorithms, we demonstrate that our algorithm produces global interpretations that are both faithful (highly accurate) and parsimonious (involve a small number of terms). Our interpretations permit easy understanding of the relative importance of features and feature interactions. Our interpretation algorithm represents a leap forward from the previous state of the art.

[Breaking Reversibility Accelerates Langevin Dynamics for Non-Convex Optimization](#)

- Xuefeng GAO, Mert Gurbuzbalaban, Lingjiong Zhu
- abstract@[open-review](#): Langevin dynamics (LD) has been proven to be a powerful technique for optimizing a non-convex objective as an efficient algorithm to find local minima while eventually visiting a global minimum on longer time-scales. LD is based on the first-order Langevin diffusion which is reversible in time. We study two variants that are based on non-reversible Langevin diffusions: the underdamped Langevin dynamics (ULD) and the Langevin dynamics with a non-symmetric drift (NLD). Adopting the techniques of Tzen et al. (2018) for LD to non-reversible diffusions, we show that for a given local minimum that is within an arbitrary distance from the initialization, with high probability, either the ULD trajectory ends up somewhere outside a small neighborhood of this local minimum within a recurrence time which depends on the smallest eigenvalue of the Hessian at the local minimum or they enter this neighborhood by the recurrence time and stay there for a potentially exponentially long escape time. The ULD algorithm improves upon the recurrence time obtained for LD in Tzen et al. (2018) with respect to the dependency on the smallest eigenvalue of the Hessian at the local minimum. Similar results and improvements are obtained for the NLD algorithm. We also show that non-reversible variants can exit the basin of attraction of a local minimum faster in discrete time when the objective has two local minima separated by a saddle point and quantify the amount of improvement. Our analysis suggests that non-reversible Langevin algorithms are more efficient to locate a local minimum as well as exploring the state space.

[Robust large-margin learning in hyperbolic space](#)

- Melanie Weber, Manzil Zaheer, Ankit Singh Rawat, Aditya K. Menon, Sanjiv Kumar
- abstract@[open-review](#): Recently, there has been a surge of interest in representation learning in hyperbolic spaces, driven by their ability to represent hierarchical data with significantly fewer dimensions than standard Euclidean spaces. However, the viability and benefits of hyperbolic spaces for downstream machine learning tasks have received less attention. In this paper, we present, to our knowledge, the first theoretical guarantees for learning a classifier in hyperbolic rather than Euclidean space. Specifically, we consider the problem of learning a large-margin classifier for data possessing a hierarchical structure. Our first contribution is a hyperbolic perceptron algorithm, which provably converges to a separating hyperplane. We then provide an algorithm to efficiently learn a large-margin hyperplane, relying on the careful injection of adversarial examples. Finally, we prove that for hierarchical data that embeds well into hyperbolic space, the low embedding dimension ensures superior guarantees when learning the classifier directly in hyperbolic space.

[Replica-Exchange Nos\\e-Hoover Dynamics for Bayesian Learning on Large Datasets](#)

- Rui Luo, Qiang Zhang, Yaodong Yang, Jun Wang
- abstract@[open-review](#): In this paper, we present a new practical method for Bayesian learning that can rapidly draw representative samples from complex posterior distributions with multiple isolated modes in the presence of mini-batch noise. This is achieved by simulating a collection of replicas in parallel with different temperatures and periodically swapping them. When evolving the replicas' states, the Nos\\e-Hoover dynamics is applied, which adaptively neutralizes the mini-batch noise. To perform proper exchanges, a new protocol is developed with a noise-aware test of acceptance, by which the detailed balance is reserved in an asymptotic way. While its efficacy on complex multimodal posteriors has been illustrated by testing over synthetic distributions, experiments with deep Bayesian neural networks on large-scale datasets have shown its significant improvements over strong baselines.

[Adversarially Robust Few-Shot Learning: A Meta-Learning Approach](#)

- Micah Goldblum, Liam Fowl, Tom Goldstein
- abstract@[open-review](#): Previous work on adversarially robust neural networks for image classification requires large training sets and computationally expensive training procedures. On the other hand, few-shot learning methods are highly vulnerable to adversarial examples. The goal of our work is to produce networks which both perform well at few-shot classification tasks and are simultaneously robust to adversarial examples. We develop an algorithm, called Adversarial Querying (AQ), for producing adversarially robust meta-learners, and we thoroughly investigate the causes for adversarial vulnerability. Moreover, our method achieves far superior robust performance on few-shot image classification tasks, such as Mini-ImageNet and CIFAR-FS, than robust transfer learning.

[Neural Anisotropy Directions](#)

- Guillermo Ortiz-Jimenez, Apostolos Modas, Seyed-Mohsen Moosavi, Pascal Frossard
- abstract@[open-review](#): In this work, we analyze the role of the network architecture in shaping the inductive bias of deep classifiers. To that end, we start by focusing on a very simple problem, i.e., classifying a class of linearly separable distributions, and show that, depending on the direction of the discriminative feature of the distribution, many state-of-the-art deep convolutional neural networks (CNNs) have a surprisingly hard time solving this

simple task. We then define as neural anisotropy directions (NADs) the vectors that encapsulate the directional inductive bias of an architecture. These vectors, which are specific for each architecture and hence act as a signature, encode the preference of a network to separate the input data based on some particular features. We provide an efficient method to identify NADs for several CNN architectures and thus reveal their directional inductive biases. Furthermore, we show that, for the CIFAR-10 dataset, NADs characterize the features used by CNNs to discriminate between different classes.

[Digraph Inception Convolutional Networks](#)

- Zekun Tong, Yuxuan Liang, Changsheng Sun, Xinkie Li, David Rosenblum, Andrew Lim
- abstract@[open-review](#): Graph Convolutional Networks (GCNs) have shown promising results in modeling graph-structured data. However, they have difficulty with processing digraphs because of two reasons: 1) transforming directed to undirected graph to guarantee the symmetry of graph Laplacian is not reasonable since it not only misleads message passing scheme to aggregate incorrect weights but also deprives the unique characteristics of digraph structure; 2) due to the fixed receptive field in each layer, GCNs fail to obtain multi-scale features that can boost their performance. In this paper, we theoretically extend spectral-based graph convolution to digraphs and derive a simplified form using personalized PageRank. Specifically, we present the Digraph Inception Convolutional Networks (DiGCN) which utilizes digraph convolution and kth-order proximity to achieve larger receptive fields and learn multi-scale features in digraphs. We empirically show that DiGCN can encode more structural information from digraphs than GCNs and help achieve better performance when generalized to other models. Moreover, experiments on various benchmarks demonstrate its superiority against the state-of-the-art methods.

[PAC-Bayesian Bound for the Conditional Value at Risk](#)

- Zakaria Mhammedi, Benjamin Guedj, Robert C. Williamson
- abstract@[open-review](#): Conditional Value at Risk (CVaR) is a ``coherent risk measure'' which generalizes expectation (reduced to a boundary parameter setting). Widely used in mathematical finance, it is garnering increasing interest in machine learning as an alternate approach to regularization, and as a means for ensuring fairness.

This paper presents a generalization bound for learning algorithms that minimize the CVaR of the empirical loss. The bound is of PAC-Bayesian type and is guaranteed to be small when the empirical CVaR is small. We achieve this by reducing the problem of estimating CVaR to that of merely estimating an expectation. This then enables us, as a by-product, to obtain concentration inequalities for CVaR even when the random variable in question is unbounded.

[Stochastic Stein Discrepancies](#)

- Jackson Gorham, Anant Raj, Lester Mackey
- abstract@[open-review](#): Stein discrepancies (SDs) monitor convergence and non-convergence in approximate inference when exact integration and sampling are intractable. However, the computation of a Stein discrepancy can be prohibitive if the Stein operator -- often a sum over likelihood terms or potentials -- is expensive to evaluate. To address this deficiency, we show that stochastic Stein discrepancies (SSDs) based on subsampled approximations of the Stein operator inherit the convergence control properties of standard SDs with probability 1. Along the way, we establish the convergence of Stein variational gradient descent (SVGD) on unbounded domains, resolving an open question of Liu (2017). In our experiments with biased Markov chain Monte Carlo (MCMC) hyperparameter tuning, approximate MCMC sampler selection, and stochastic SVGD, SSDs deliver comparable inferences to standard SDs with orders of magnitude fewer likelihood evaluations.

[On the Role of Sparsity and DAG Constraints for Learning Linear DAGs](#)

- Ignavier Ng, AmirEmad Ghassami, Kun Zhang
- abstract@[open-review](#): Learning graphical structure based on Directed Acyclic Graphs (DAGs) is a challenging problem, partly owing to the large search space of possible graphs. A recent line of work formulates the structure learning problem as a continuous constrained optimization task using the least squares objective and an algebraic characterization of DAGs. However, the formulation requires a hard DAG constraint and may lead to optimization difficulties. In this paper, we study the asymptotic role of the sparsity and DAG constraints for learning DAG models in the linear Gaussian and non-Gaussian cases, and investigate their usefulness in the finite sample regime. Based on the theoretical results, we formulate a likelihood-based score function, and show that one only has to apply soft sparsity and DAG constraints to learn a DAG equivalent to the ground truth DAG. This leads to an unconstrained optimization problem that is much easier to solve. Using gradient-based optimization and GPU acceleration, our procedure can easily handle thousands of nodes while retaining a high accuracy. Extensive experiments validate the effectiveness of our proposed method and show that the DAG-penalized likelihood objective is indeed favorable over the least squares one with the hard DAG constraint.

[Cream of the Crop: Distilling Prioritized Paths For One-Shot Neural Architecture Search](#)

- Houwen Peng, Hao Du, Hongyuan Yu, QI LI, Jing Liao, Jianlong Fu
- abstract@[open-review](#): One-shot weight sharing methods have recently drawn great attention in neural architecture search due to high efficiency and competitive performance. However, weight sharing across models has an inherent deficiency, i.e., insufficient training of subnetworks in the hypernetwork. To alleviate this problem, we present a simple yet effective architecture distillation method. The central idea is that subnetworks can learn collaboratively and teach each other throughout the training process, aiming to boost the convergence of individual models. We introduce the concept of prioritized path, which refers to the architecture candidates exhibiting superior performance during training. Distilling knowledge from the prioritized paths is able to boost the training of subnetworks. Since the prioritized paths are changed on the fly depending on their performance and complexity, the final obtained paths are the cream of the crop. We directly select the most promising one from the prioritized paths as the final architecture, without using other complex search methods, such as reinforcement learning or evolution algorithms. The experiments on ImageNet verify such path distillation method can improve the convergence ratio and performance of the hypernetwork, as well as boosting the training of subnetworks. The discovered architectures achieve superior performance compared to the recent MobileNetV3 and EfficientNet families under aligned settings. Moreover, the experiments on object detection and more challenging search space show the generality and robustness of the proposed method. Code and models are available at \url{https://github.com/neurips-20/cream.git}.

[Fair Multiple Decision Making Through Soft Interventions](#)

- Yaowei Hu, Yongkai Wu, Lu Zhang, Xintao Wu
- abstract@[open-review](#): Previous research in fair classification mostly focuses on a single decision model. In reality, there usually exist multiple decision models within a system and all of which may contain a certain amount of discrimination. Such realistic scenarios introduce new challenges to fair classification: since discrimination may be transmitted from upstream models to downstream models, building decision models separately without taking upstream models into consideration cannot guarantee to achieve fairness. In this paper, we propose an approach that learns multiple classifiers and achieves fairness for all of them simultaneously, by treating each decision model as a soft intervention and inferring the post-intervention distributions to formulate the loss function as well as the fairness constraints. We adopt surrogate functions to smooth the loss function and constraints, and theoretically show that the excess risk of the proposed loss function can be bounded in a form that is the same as that for traditional surrogated loss functions. Experiments using both synthetic and real-world datasets show the effectiveness of our approach.

[Representation Learning for Integrating Multi-domain Outcomes to Optimize Individualized Treatment](#)

- Yuan Chen, Donglin Zeng, Tianchen Xu, Yuanjia Wang
- abstract@[open-review](#): For mental disorders, patients' underlying mental states are non-observed latent constructs which have to be inferred from observed multi-domain measurements such as diagnostic symptoms and patient functioning scores. Additionally, substantial heterogeneity in the disease diagnosis between patients needs to be addressed for optimizing individualized treatment policy in order to achieve precision medicine. To address these challenges, we propose an integrated learning framework that can simultaneously learn patients' underlying mental states and recommend optimal treatments for each individual. This learning framework is based on the measurement theory in psychiatry for modeling multiple disease diagnostic measures as arising from the underlying causes (true mental states). It allows incorporation of the multivariate pre- and post-treatment outcomes as well as biological measures while preserving the invariant structure for representing patients' latent mental states. A multi-layer neural network is used to allow complex treatment effect heterogeneity. Optimal treatment policy can be inferred for future patients by comparing their potential mental states under different treatments given the observed multi-domain pre-treatment measurements. Experiments on simulated data and a real-world clinical trial data show that the learned treatment policies compare favorably to alternative methods on heterogeneous treatment effects, and have broad utilities which lead to better patient outcomes on multiple domains.

[Learning to Play No-Press Diplomacy with Best Response Policy Iteration](#)

- Thomas Anthony, Tom Eccles, Andrea Tacchetti, JÃ©rÃ©mias KramÃ©r, Ian Gemp, Thomas Hudson, Nicolas Porcel, Marc Lanctot, Julien Perolat, Richard Everett, Satinder Singh, Thore Graepel, Yoram Bachrach
- abstract@[open-review](#): Recent advances in deep reinforcement learning (RL) have led to considerable progress in many 2-player zero-sum games, such as Go, Poker and Starcraft. The purely adversarial nature of such games allows for conceptually simple and principled application of RL methods. However real-world settings are many-agent, and agent interactions are complex mixtures of common-interest and competitive aspects. We consider Diplomacy, a 7-player board game designed to accentuate dilemmas resulting from many-agent interactions. It also features a large combinatorial action space and simultaneous moves, which are challenging for RL algorithms. We propose a simple yet effective approximate best response operator, designed to handle large combinatorial action spaces and simultaneous moves. We also introduce a family of policy iteration methods that approximate fictitious play. With these methods, we successfully apply RL to Diplomacy: we show that our agents convincingly outperform the previous state-of-the-art, and game theoretic equilibrium analysis shows that the new process yields consistent improvements.

[Inverse Learning of Symmetries](#)

- Mario Wieser, Sonali Parbhoo, Aleksander Wieczorek, Volker Roth
- abstract@[open-review](#): Symmetry transformations induce invariances and are a crucial building block of modern machine learning algorithms. In many complex domains, such as the chemical space, invariances can be observed, yet the corresponding symmetry transformation cannot be formulated analytically. We propose to learn the symmetry transformation with a model consisting of two latent subspaces, where the first subspace captures the target and the second subspace the remaining invariant information. Our approach is based on the deep information bottleneck in combination with a continuous mutual information regulariser. Unlike previous methods, we focus on the challenging task of minimising mutual information in continuous domains. To this end, we base the calculation of mutual information on correlation matrices in combination with a bijective variable transformation. Extensive experiments demonstrate that our model outperforms state-of-the-art methods on artificial and molecular datasets.

[DiffGCN: Graph Convolutional Networks via Differential Operators and Algebraic Multigrid Pooling](#)

- Moshe Eliasof, Eran Treister
- abstract@[open-review](#): Graph Convolutional Networks (GCNs) have shown to be effective in handling unordered data like point clouds and meshes. In this work we propose novel approaches for graph convolution, pooling and unpooling, inspired from finite differences and algebraic multigrid frameworks. We form a parameterized convolution kernel based on discretized differential operators, leveraging the graph mass, gradient and Laplacian. This way, the parameterization does not depend on the graph structure, only on the meaning of the network convolutions as differential operators. To allow hierarchical representations of the input, we propose pooling and unpooling operations that are based on algebraic multigrid methods, which are mainly used to solve partial differential equations on unstructured grids. To motivate and explain our method, we compare it to standard convolutional neural networks, and show their similarities and relations in the case of a regular grid. Our proposed method is demonstrated in various experiments like classification and part-segmentation, achieving on par or better than state of the art results. We also analyze the computational cost of our method compared to other GCNs.

[Distributed Newton Can Communicate Less and Resist Byzantine Workers](#)

- Avishek Ghosh, Raj Kumar Maity, Arya Mazumdar
- abstract@[open-review](#): We develop a distributed second order optimization algorithm that is communication-efficient as well as robust against Byzantine failures of the worker machines. We propose an iterative approximate Newton-type algorithm, where the worker machines communicate \emph{only once} per iteration with the central machine. This is in sharp contrast with the state-of-the-art distributed second order algorithms like GIANT \cite{giant}, DINGO\cite{dingo}, where the worker machines send (functions of) local gradient and Hessian sequentially; thus ending up communicating twice with the central machine per iteration. Furthermore, we employ a simple norm based thresholding rule to filter-out the Byzantine worker machines. We establish the linear-quadratic rate of convergence of our proposed algorithm and establish that the communication savings and Byzantine resilience attributes only correspond to a small statistical error rate for arbitrary convex loss functions. To the best of our knowledge, this is the first work that addresses the issue of Byzantine resilience in second order distributed optimization. Furthermore, we validate our theoretical results with extensive experiments on synthetically generated and benchmark LIBSVM \cite{libsvm} data-set and demonstrate convergence guarantees.

[Efficient Nonmyopic Bayesian Optimization via One-Shot Multi-Step Trees](#)

- Shali Jiang, Daniel Jiang, Maximilian Balandat, Brian Karrer, Jacob Gardner, Roman Garnett
- abstract@[open-review](#): Bayesian optimization is a sequential decision making framework for optimizing expensive-to-evaluate black-box functions. Computing a full lookahead policy amounts to solving a highly intractable stochastic dynamic program. Myopic approaches, such as expected improvement, are often adopted in practice, but they ignore the long-term impact of the immediate decision. Existing nonmyopic approaches are mostly heuristic and/or computationally expensive. In this paper, we provide the first efficient implementation of general multi-step lookahead Bayesian optimization, formulated as a sequence of nested optimization problems within a multi-step scenario tree. Instead of solving these problems in a nested way, we equivalently optimize all decision variables in the full tree jointly, in a "one-shot" fashion. Combining this with an efficient method for implementing multi-step Gaussian process "fantasization," we demonstrate that multi-step expected improvement is computationally tractable and exhibits performance superior to existing methods on a wide range of benchmarks.

[Effective Diversity in Population Based Reinforcement Learning](#)

- Jack Parker-Holder, Aldo Pacchiano, Krzysztof M. Choromanski, Stephen J. Roberts

- abstract@[open-review](#): Exploration is a key problem in reinforcement learning, since agents can only learn from data they acquire in the environment. With that in mind, maintaining a population of agents is an attractive method, as it allows data to be collected with a diverse set of behaviors. This behavioral diversity is often boosted via multi-objective loss functions. However, those approaches typically leverage mean field updates based on pairwise distances, which makes them susceptible to cycling behaviors and increased redundancy. In addition, explicitly boosting diversity often has a detrimental impact on optimizing already fruitful behaviors for rewards. As such, the reward-diversity trade off typically relies on heuristics. Finally, such methods require behavioral representations, often handcrafted and domain specific. In this paper, we introduce an approach to optimize all members of a population simultaneously. Rather than using pairwise distance, we measure the volume of the entire population in a behavioral manifold, defined by task-agnostic behavioral embeddings. In addition, our algorithm Diversity via Determinants (DvD), adapts the degree of diversity during training using online learning techniques. We introduce both evolutionary and gradient-based instantiations of DvD and show they effectively improve exploration without reducing performance when better exploration is not required.

[Elastic-InfoGAN: Unsupervised Disentangled Representation Learning in Class-Imbalanced Data](#)

- Utkarsh Ojha, Krishna Kumar Singh, Cho-Jui Hsieh, Yong Jae Lee
- abstract@[open-review](#): We propose a novel unsupervised generative model that learns to disentangle object identity from other low-level aspects in class-imbalanced data. We first investigate the issues surrounding the assumptions about uniformity made by InfoGAN, and demonstrate its ineffectiveness to properly disentangle object identity in imbalanced data. Our key idea is to make the discovery of the discrete latent factor of variation invariant to identity-preserving transformations in real images, and use that as a signal to learn the appropriate latent distribution representing object identity. Experiments on both artificial (MNIST, 3D cars, 3D chairs, ShapeNet) and real-world (YouTube-Faces) imbalanced datasets demonstrate the effectiveness of our method in disentangling object identity as a latent factor of variation.

[Direct Policy Gradients: Direct Optimization of Policies in Discrete Action Spaces](#)

- Guy Lorberbom, Chris J. Maddison, Nicolas Heess, Tamir Hazan, Daniel Tarlow
- abstract@[open-review](#): Direct optimization (McAllester et al., 2010; Song et al., 2016) is an appealing framework that replaces integration with optimization of a random objective for approximating gradients in models with discrete random variables (Lorberbom et al., 2018). A* sampling (Maddison et al., 2014) is a framework for optimizing such random objectives over large spaces. We show how to combine these techniques to yield a reinforcement learning algorithm that approximates a policy gradient by finding trajectories that optimize a random objective. We call the resulting algorithms \emph{direct policy gradient} (DirPG) algorithms. A main benefit of DirPG algorithms is that they allow the insertion of domain knowledge in the form of upper bounds on return-to-go at training time, like is used in heuristic search, while still directly computing a policy gradient. We further analyze their properties, showing there are cases where DirPG has an exponentially larger probability of sampling informative gradients compared to REINFORCE. We also show that there is a built-in variance reduction technique and that a parameter that was previously viewed as a numerical approximation can be interpreted as controlling risk sensitivity. Empirically, we evaluate the effect of key degrees of freedom and show that the algorithm performs well in illustrative domains compared to baselines.

[Hybrid Models for Learning to Branch](#)

- Prateek Gupta, Maxime Gasse, Elias Khalil, Pawan Mudigonda, Andrea Lodi, Yoshua Bengio
- abstract@[open-review](#): A recent Graph Neural Network (GNN) approach for learning to branch has been shown to successfully reduce the running time of branch-and-bound algorithms for Mixed Integer Linear Programming (MILP). While the GNN relies on a GPU for inference, MILP solvers are purely CPU-based. This severely limits its application as many practitioners may not have access to high-end GPUs. In this work, we ask two key questions. First, in a more realistic setting where only a CPU is available, is the GNN model still competitive? Second, can we devise an alternate computationally inexpensive model that retains the predictive power of the GNN architecture? We answer the first question in the negative, and address the second question by proposing a new hybrid architecture for efficient branching on CPU machines. The proposed architecture combines the expressive power of GNNs with computationally inexpensive multi-layer perceptrons (MLP) for branching. We evaluate our methods on four classes of MILP problems, and show that they lead to up to 26% reduction in solver running time compared to state-of-the-art methods without a GPU, while extrapolating to harder problems than it was trained on. The code for this project is publicly available at <https://github.com/pg2455/Hybrid-learn2branch>.

[WoodFisher: Efficient Second-Order Approximation for Neural Network Compression](#)

- Sidak Pal Singh, Dan Alistarh
- abstract@[open-review](#): Second-order information, in the form of Hessian- or Inverse-Hessian-vector products, is a fundamental tool for solving optimization problems. Recently, there has been significant interest in utilizing this information in the context of deep neural networks; however, relatively little is known about the quality of existing approximations in this context. Our work considers this question, examines the accuracy of existing approaches, and proposes a method called WoodFisher to compute a faithful and efficient estimate of the inverse Hessian. Our main application is to neural network compression, where we build on the classic Optimal Brain Damage/Surgeon framework. We demonstrate that WoodFisher significantly outperforms popular state-of-the-art methods for one-shot pruning. Further, even when iterative, gradual pruning is allowed, our method results in a gain in test accuracy over the state-of-the-art approaches for popular image classification datasets such as ImageNet ILSVRC. Further, we show how our method can be extended to take into account first-order information, and illustrate its ability to automatically set layer-wise pruning thresholds, or perform compression in the limited-data regime.

[Bi-level Score Matching for Learning Energy-based Latent Variable Models](#)

- Fan Bao, Chongxuan LI, Kun Xu, Hang Su, Jun Zhu, Bo Zhang
- abstract@[open-review](#): Score matching (SM) provides a compelling approach to learn energy-based models (EBMs) by avoiding the calculation of partition function. However, it remains largely open to learn energy-based latent variable models (EBLVMs), except some special cases. This paper presents a bi-level score matching (BiSM) method to learn EBLVMs with general structures by reformulating SM as a bi-level optimization problem. The higher level introduces a variational posterior of the latent variables and optimizes a modified SM objective, and the lower level optimizes the variational posterior to fit the true posterior. To solve BiSM efficiently, we develop a stochastic optimization algorithm with gradient unrolling. Theoretically, we analyze the consistency of BiSM and the convergence of the stochastic algorithm. Empirically, we show the promise of BiSM in Gaussian restricted Boltzmann machines and highly nonstructural EBLVMs parameterized by deep convolutional neural networks. BiSM is comparable to the widely adopted contrastive divergence and SM methods when they are applicable; and can learn complex EBLVMs with intractable posteriors to generate natural images.

[Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding](#)

- Zhu Zhang, Zhou Zhao, Zhijie Lin, jieming zhu, Xiuqiang He
- abstract@[open-review](#): Weakly-supervised vision-language grounding aims to localize a target moment in a video or a specific region in an image according to the given sentence query, where only video-level or image-level sentence annotations are provided during training. Most existing approaches employ the MIL-based or reconstruction-based paradigms for the WSLVG task, but the former heavily depends on the quality of randomly-selected negative samples and the latter cannot directly optimize the visual-textual alignment score. In this paper, we propose a novel Counterfactual Contrastive Learning (CCL) to develop sufficient contrastive training between counterfactual positive and negative results, which are based on robust and destructive counterfactual transformations. Concretely, we design three counterfactual transformation strategies from the feature-, interaction- and relation-level,

where the feature-level method damages the visual features of selected proposals, interaction-level approach confuses the vision-language interaction and relation-level strategy destroys the context clues in proposal relationships. Extensive experiments on five vision-language grounding datasets verify the effectiveness of our CCL paradigm.

[Decision trees as partitioning machines to characterize their generalization properties](#)

- Jean-Samuel Leboeuf, FrÃ©dÃ©ric LeBlanc, Mario Marchand
- abstract@[open-review](#): Decision trees are popular machine learning models that are simple to build and easy to interpret. Even though algorithms to learn decision trees date back to almost 50 years, key properties affecting their generalization error are still weakly bounded. Hence, we revisit binary decision trees on real-valued features from the perspective of partitions of the data. We introduce the notion of partitioning function, and we relate it to the growth function and to the VC dimension. Using this new concept, we are able to find the exact VC dimension of decision stumps, which is given by the largest integer d such that $\ell \geq \binom{d}{\lfloor \frac{d}{2} \rfloor}$, where ℓ is the number of real-valued features. We provide a recursive expression to bound the partitioning functions, resulting in an upper bound on the growth function of any decision tree structure. This allows us to show that the VC dimension of a binary tree structure with N internal nodes is of order $N \log(N)$. Finally, we elaborate a pruning algorithm based on these results that performs better than the CART algorithm on a number of datasets, with the advantage that no cross-validation is required.

[Learning to Prove Theorems by Learning to Generate Theorems](#)

- Mingzhe Wang, Jia Deng
- abstract@[open-review](#): We consider the task of automated theorem proving, a key AI task. Deep learning has shown promise for training theorem provers, but there are limited human-written theorems and proofs available for supervised learning. To address this limitation, we propose to learn a neural generator that automatically synthesizes theorems and proofs for the purpose of training a theorem prover. Experiments on real-world tasks demonstrate that synthetic data from our approach improves the theorem prover and advances the state of the art of automated theorem proving in Metamath.

[3D Self-Supervised Methods for Medical Imaging](#)

- Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, Christoph Lippert
- abstract@[open-review](#): Self-supervised learning methods have witnessed a recent surge of interest after proving successful in multiple application fields. In this work, we leverage these techniques, and we propose 3D versions for five different self-supervised methods, in the form of proxy tasks. Our methods facilitate neural network feature learning from unlabeled 3D images, aiming to reduce the required cost for expert annotation. The developed algorithms are 3D Contrastive Predictive Coding, 3D Rotation prediction, 3D Jigsaw puzzles, Relative 3D patch location, and 3D Exemplar networks. Our experiments show that pretraining models with our 3D tasks yields more powerful semantic representations, and enables solving downstream tasks more accurately and efficiently, compared to training the models from scratch and to pretraining them on 2D slices. We demonstrate the effectiveness of our methods on three downstream tasks from the medical imaging domain: i) Brain Tumor Segmentation from 3D MRI, ii) Pancreas Tumor Segmentation from 3D CT, and iii) Diabetic Retinopathy Detection from 2D Fundus images. In each task, we assess the gains in data-efficiency, performance, and speed of convergence. Interestingly, we also find gains when transferring the learned representations, by our methods, from a large unlabeled 3D corpus to a small downstream-specific dataset. We achieve results competitive to state-of-the-art solutions at a fraction of the computational expense. We publish our implementations for the developed algorithms (both 3D and 2D versions) as an open-source library, in an effort to allow other researchers to apply and extend our methods on their datasets.

[Bayesian filtering unifies adaptive and non-adaptive neural network optimization methods](#)

- Laurence Aitchison
- abstract@[open-review](#): We formulate the problem of neural network optimization as Bayesian filtering, where the observations are backpropagated gradients. While neural network optimization has previously been studied using natural gradient methods which are closely related to Bayesian inference, they were unable to recover standard optimizers such as Adam and RMSprop with a root-mean-square gradient normalizer, instead getting a mean-square normalizer. To recover the root-mean-square normalizer, we find it necessary to account for the temporal dynamics of all the other parameters as they are optimized. The resulting optimizer, AdaBayes, adaptively transitions between SGD-like and Adam-like behaviour, automatically recovers AdamW, a state of the art variant of Adam with decoupled weight decay, and has generalisation performance competitive with SGD.

[Worst-Case Analysis for Randomly Collected Data](#)

- Justin Chen, Gregory Valiant, Paul Valiant
- abstract@[open-review](#): We introduce a framework for statistical estimation that leverages knowledge of how samples are collected but makes no distributional assumptions on the data values. Specifically, we consider a population of elements $[n] = \{1, \dots, n\}$ with corresponding data values x_1, \dots, x_n . We observe the values for a "sample" set $A \subset [n]$ and wish to estimate some statistic of the values for a "target" set $B \subset [n]$ where B could be the entire set. Crucially, we assume that the sets A and B are drawn according to some known distribution P over pairs of subsets of $[n]$. A given estimation algorithm is evaluated based on its "worst-case, expected error" where the expectation is with respect to the distribution P from which the sample A and target sets B are drawn, and the worst-case is with respect to the data values x_1, \dots, x_n . Within this framework, we give an efficient algorithm for estimating the target mean that returns a weighted combination of the sample values—where the weights are functions of the distribution P and the sample and target sets A, B —and show that the worst-case expected error achieved by this algorithm is at most a multiplicative $\pi/2$ factor worse than the optimal of such algorithms. The algorithm and proof leverage a surprising connection to the Grothendieck problem. We also extend these results to the linear regression setting where each datapoint is not a scalar but a labeled vector (x_i, y_i) . This framework, which makes no distributional assumptions on the data values but rather relies on knowledge of the data collection process via the distribution P , is a significant departure from the typical statistical estimation framework and introduces a uniform analysis for the many natural settings where membership in a sample may be correlated with data values, such as when individuals are recruited into a sample through their social networks as in "snowball/chain" sampling or when samples have chronological structure as in "selective prediction".

[Truthful Data Acquisition via Peer Prediction](#)

- Yiling Chen, Yiheng Shen, Shuran Zheng
- abstract@[open-review](#): We consider the problem of purchasing data for machine learning or statistical estimation. The data analyst has a budget to purchase datasets from multiple data providers. She does not have any test data that can be used to evaluate the collected data and can assign payments to data providers solely based on the collected datasets. We consider the problem in the standard Bayesian paradigm and in two settings: (1) data are only collected once; (2) data are collected repeatedly and each day's data are drawn independently from the same distribution. For both settings, our mechanisms guarantee that truthfully reporting one's dataset is always an equilibrium by adopting techniques from peer prediction: pay each provider the mutual information between his reported data and other providers' reported data. Depending on the data distribution, the mechanisms can also discourage misreports that would lead to inaccurate predictions. Our mechanisms also guarantee individual rationality and budget feasibility for certain underlying distributions in the first setting and for all distributions in the second setting.

[Learning Robust Decision Policies from Observational Data](#)

- Muhammad Osama, Dave Zachariah, Peter Stoica
- abstract@[open-review](#): We address the problem of learning a decision policy from observational data of past decisions in contexts with features and associated outcomes. The past policy maybe unknown and in safety-critical applications, such as medical decision support, it is of interest to learn robust policies that reduce the risk of outcomes with high costs. In this paper, we develop a method for learning policies that reduce tails of the cost distribution at a specified level and, moreover, provide a statistically valid bound on the cost of each decision. These properties are valid under finite samples -- even in scenarios with uneven or no overlap between features for different decisions in the observed data -- by building on recent results in conformal prediction. The performance and statistical properties of the proposed method are illustrated using both real and synthetic data.

[Byzantine Resilient Distributed Multi-Task Learning](#)

- Jian Li, Waseem Abbas, Xenofon Koutsoukos
- abstract@[open-review](#): Distributed multi-task learning provides significant advantages in multi-agent networks with heterogeneous data sources where agents aim to learn distinct but correlated models simultaneously. However, distributed algorithms for learning relatedness among tasks are not resilient in the presence of Byzantine agents. In this paper, we present an approach for Byzantine resilient distributed multi-task learning. We propose an efficient online weight assignment rule by measuring the accumulated loss using an agent's data and its neighbors' models. A small accumulated loss indicates a large similarity between the two tasks. In order to ensure the Byzantine resilience of the aggregation at a normal agent, we introduce a step for filtering out larger losses. We analyze the approach for convex models and show that normal agents converge resiliently towards their true targets. Further, an agent's learning performance using the proposed weight assignment rule is guaranteed to be at least as good as in the non-cooperative case as measured by the expected regret. Finally, we demonstrate the approach using three case studies, including regression and classification problems, and show that our method exhibits good empirical performance for non-convex models, such as convolutional neural networks.

[Reinforcement Learning in Factored MDPs: Oracle-Efficient Algorithms and Tighter Regret Bounds for the Non-Episodic Setting](#)

- Ziping Xu, Ambuj Tewari
- abstract@[open-review](#): We study reinforcement learning in non-episodic factored Markov decision processes (FMDPs). We propose two near-optimal and oracle-efficient algorithms for FMDPs. Assuming oracle access to an FMDP planner, they enjoy a Bayesian and a frequentist regret bound respectively, both of which reduce to the near-optimal bound $\$O(DS\sqrt{AT})\$$ for standard non-factored MDPs. We propose a tighter connectivity measure, factored span, for FMDPs and prove a lower bound that depends on the factored span rather than the diameter D . In order to decrease the gap between lower and upper bounds, we propose an adaptation of the REGAL.C algorithm whose regret bound depends on the factored span. Our oracle-efficient algorithms outperform previously proposed near-optimal algorithms on computer network administration simulations.

[Improving model calibration with accuracy versus uncertainty optimization](#)

- Ranganath Krishnan, Omesh Tickoo
- abstract@[open-review](#): Obtaining reliable and accurate quantification of uncertainty estimates from deep neural networks is important in safety-critical applications. A well-calibrated model should be accurate when it is certain about its prediction and indicate high uncertainty when it is likely to be inaccurate. Uncertainty calibration is a challenging problem as there is no ground truth available for uncertainty estimates. We propose an optimization method that leverages the relationship between accuracy and uncertainty as an anchor for uncertainty calibration. We introduce a differentiable accuracy versus uncertainty calibration (AvUC) loss function that allows a model to learn to provide well-calibrated uncertainties, in addition to improved accuracy. We also demonstrate the same methodology can be extended to post-hoc uncertainty calibration on pretrained models. We illustrate our approach with mean-field stochastic variational inference and compare with state-of-the-art methods. Extensive experiments demonstrate our approach yields better model calibration than existing methods on large-scale image classification tasks under distributional shift.

[The Convolution Exponential and Generalized Sylvester Flows](#)

- Emiel Hoogeboom, Victor Garcia Satorras, Jakub Tomczak, Max Welling
- abstract@[open-review](#): This paper introduces a new method to build linear flows, by taking the exponential of a linear transformation. This linear transformation does not need to be invertible itself, and the exponential has the following desirable properties: it is guaranteed to be invertible, its inverse is straightforward to compute and the log Jacobian determinant is equal to the trace of the linear transformation. An important insight is that the exponential can be computed implicitly, which allows the use of convolutional layers. Using this insight, we develop new invertible transformations named convolution exponentials and graph convolution exponentials, which retain the equivariance of their underlying transformations. In addition, we generalize Sylvester Flows and propose Convolutional Sylvester Flows which are based on the generalization and the convolution exponential as basis change. Empirically, we show that the convolution exponential outperforms other linear transformations in generative flows on CIFAR10 and the graph convolution exponential improves the performance of graph normalizing flows. In addition, we show that Convolutional Sylvester Flows improve performance over residual flows as a generative flow model measured in log-likelihood.

[An Improved Analysis of Stochastic Gradient Descent with Momentum](#)

- Yanli Liu, Yuan Gao, Wotao Yin
- abstract@[open-review](#): SGD with momentum (SGDM) has been widely applied in many machine learning tasks, and it is often applied with dynamic stepsizes and momentum weights tuned in a stagewise manner. Despite of its empirical advantage over SGD, the role of momentum is still unclear in general since previous analyses on SGDM either provide worse convergence bounds than those of SGD, or assume Lipschitz or quadratic objectives, which fail to hold in practice. Furthermore, the role of dynamic parameters has not been addressed. In this work, we show that SGDM converges as fast as SGD for smooth objectives under both strongly convex and nonconvex settings. We also prove that multistage strategy is beneficial for SGDM compared to using fixed parameters. Finally, we verify these theoretical claims by numerical experiments.

[Precise expressions for random projections: Low-rank approximation and randomized Newton](#)

- Michal Derezhinski, Feynman T. Liang, Zhenyu Liao, Michael W. Mahoney
- abstract@[open-review](#): It is often desirable to reduce the dimensionality of a large dataset by projecting it onto a low-dimensional subspace. Matrix sketching has emerged as a powerful technique for performing such dimensionality reduction very efficiently. Even though there is an extensive literature on the worst-case performance of sketching, existing guarantees are typically very different from what is observed in practice. We exploit recent developments in the spectral analysis of random matrices to develop novel techniques that provide provably accurate expressions for the expected value of random projection matrices obtained via sketching. These expressions can be used to characterize the performance of dimensionality reduction in a variety of common machine learning tasks, ranging from low-rank approximation to iterative stochastic optimization. Our results apply to several popular sketching methods, including Gaussian and Rademacher sketches, and they enable precise analysis of these methods in terms of spectral properties of the data. Empirical results show that the expressions we derive reflect the practical performance of these sketching methods, down to lower-order effects and even constant factors.

[The MAGICAL Benchmark for Robust Imitation](#)

- Sam Toyer, Rohin Shah, Andrew Critch, Stuart Russell
- abstract@[open-review](#): Imitation Learning (IL) algorithms are typically evaluated in the same environment that was used to create demonstrations. This rewards precise reproduction of demonstrations in one particular environment, but provides little information about how robustly an algorithm can generalise the demonstrator's intent to substantially different deployment settings. This paper presents the MAGICAL benchmark suite, which permits systematic evaluation of generalisation by quantifying robustness to different kinds of distribution shift that an IL algorithm is likely to encounter in practice. Using the MAGICAL suite, we confirm that existing IL algorithms overfit significantly to the context in which demonstrations are provided. We also show that standard methods for reducing overfitting are effective at creating narrow perceptual invariances, but are not sufficient to enable transfer to contexts that require substantially different behaviour, which suggests that new approaches will be needed in order to robustly generalise demonstrator intent. Code and data for the MAGICAL suite is available at <https://github.com/qxcv/magical/>

[X-CAL: Explicit Calibration for Survival Analysis](#)

- Mark Goldstein, Xintian Han, Aahlad Puli, Adler Perotte, Rajesh Ranganath
- abstract@[open-review](#): Survival analysis models the distribution of time until an event of interest, such as discharge from the hospital or admission to the ICU. When a model's predicted number of events within any time interval is similar to the observed number, it is called well-calibrated. A survival model's calibration can be measured using, for instance, distributional calibration (D-CALIBRATION) [Haider et al., 2020] which computes the squared difference between the observed and predicted number of events within different time intervals. Classically, calibration is addressed in post-training analysis. We develop explicit calibration (X-CAL), which turns D-CALIBRATION into a differentiable objective that can be used in survival modeling alongside maximum likelihood estimation and other objectives. X-CAL allows us to directly optimize calibration and strike a desired trade-off between predictive power and calibration. In our experiments, we fit a variety of shallow and deep models on simulated data, a survival dataset based on MNIST, on length-of-stay prediction using MIMIC-III data, and on brain cancer data from The Cancer Genome Atlas. We show that the models we study can be miscalibrated. We give experimental evidence on these datasets that X-CAL improves D-CALIBRATION without a large decrease in concordance or likelihood.

[Decentralized Accelerated Proximal Gradient Descent](#)

- Haishan Ye, Ziang Zhou, Luo Luo, Tong Zhang
- abstract@[open-review](#): Decentralized optimization has wide applications in machine learning, signal processing, and control. In this paper, we study the decentralized composite optimization problem with a non-smooth regularization term. Many proximal gradient based decentralized algorithms have been proposed in the past. However, these algorithms do not achieve near optimal computational complexity and communication complexity. In this paper, we propose a new method which establishes the optimal computational complexity and a near optimal communication complexity. Our empirical study shows that the proposed algorithm outperforms existing state-of-the-art algorithms.

[Making Non-Stochastic Control \(Almost\) as Easy as Stochastic](#)

- Max Simchowitz
- abstract@[open-review](#): Recent literature has made much progress in understanding \emph{online LQR}: a modern learning-theoretic take on the classical control problem where a learner attempts to optimally control an unknown linear dynamical system with fully observed state, perturbed by i.i.d. Gaussian noise. \iftoggle{nips}{The}{It} is now understood that the optimal regret over time horizon \$T\$ against the optimal control law scales as \$\widetilde{\|\Theta\|}\sqrt{T}\$. In this paper, we show that the same regret rate (against a suitable benchmark) is attainable even in the considerably more general non-stochastic control model, where the system is driven by \emph{arbitrary adversarial} noise \citep{agarwal2019online}. We attain the optimal \$\widetilde{\mathcal{O}}(\sqrt{T})\$ regret when the dynamics are unknown to the learner, and \$\mathcal{O}(\log T)\$ regret when known, provided that the cost functions are strongly convex (as in LQR). Our algorithm is based on a novel variant of online Newton step \citep{hazan2007logarithmic}, which adapts to the geometry induced by adversarial disturbances, and our analysis hinges on generic regret bounds for certain structured losses in the OCO-with-memory framework \citep{anava2015online}.

[BERT Loses Patience: Fast and Robust Inference with Early Exit](#)

- Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, Furu Wei
- abstract@[open-review](#): In this paper, we propose Patience-based Early Exit, a straightforward yet effective inference method that can be used as a plug-and-play technique to simultaneously improve the efficiency and robustness of a pretrained language model (PLM). To achieve this, our approach couples an internal-classifier with each layer of a PLM and dynamically stops inference when the intermediate predictions of the internal classifiers do not change for a pre-defined number of steps. Our approach improves inference efficiency as it allows the model to make a prediction with fewer layers. Meanwhile, experimental results with an ALBERT model show that our method can improve the accuracy and robustness of the model by preventing it from overthinking and exploiting multiple classifiers for prediction, yielding a better accuracy-speed trade-off compared to existing early exit methods.

[Optimal and Practical Algorithms for Smooth and Strongly Convex Decentralized Optimization](#)

- Dmitry Kovalev, Adil Salim, Peter Richtarik
- abstract@[open-review](#): We consider the task of decentralized minimization of the sum of smooth strongly convex functions stored across the nodes of a network. For this problem, lower bounds on the number of gradient computations and the number of communication rounds required to achieve \$\varepsilon\$ accuracy have recently been proven. We propose two new algorithms for this decentralized optimization problem and equip them with complexity guarantees. We show that our first method is optimal both in terms of the number of communication rounds and in terms of the number of gradient computations. Unlike existing optimal algorithms, our algorithm does not rely on the expensive evaluation of dual gradients. Our second algorithm is optimal in terms of the number of communication rounds, without a logarithmic factor. Our approach relies on viewing the two proposed algorithms as accelerated variants of the Forward Backward algorithm to solve monotone inclusions associated with the decentralized optimization problem. We also verify the efficacy of our methods against state-of-the-art algorithms through numerical experiments.

[BAIL: Best-Action Imitation Learning for Batch Deep Reinforcement Learning](#)

- Xinyue Chen, Zijian Zhou, Zheng Wang, Che Wang, Yanqiu Wu, Keith Ross
- abstract@[open-review](#): There has recently been a surge in research in batch Deep Reinforcement Learning (DRL), which aims for learning a high-performing policy from a given dataset without additional interactions with the environment. We propose a new algorithm, Best-Action Imitation Learning (BAIL), which strives for both simplicity and performance. BAIL learns a V function, uses the V function to select actions it believes to be high-performing, and then uses those actions to train a policy network using imitation learning. For the MuJoCo benchmark, we provide a comprehensive experimental study of BAIL, comparing its performance to four other batch Q-learning and imitation-learning schemes for a large variety of batch datasets. Our experiments show that BAIL's performance is much higher than the other schemes, and is also computationally much faster than the batch Q-learning schemes.

[Regularizing Towards Permutation Invariance In Recurrent Models](#)

- Edo Cohen-Karlik, Avichai Ben David, Amir Globerson
- abstract@[open-review](#): In many machine learning problems the output should not depend on the order of the inputs. Such ``permutation invariant'' functions have been studied extensively recently. Here we argue that temporal architectures such as RNNs are highly relevant for such problems, despite the inherent dependence of RNNs on order. We show that RNNs can be regularized towards permutation invariance, and that this can result in compact models, as compared to non-recursive architectures. Existing solutions (e.g., DeepSets) mostly suggest restricting the learning problem to hypothesis classes which are permutation invariant by design. Our approach of enforcing permutation invariance via regularization gives rise to learning functions which are "semi permutation invariant", e.g. invariant to some permutations and not to others. Our approach relies on a novel form of stochastic regularization. We demonstrate that our method is beneficial compared to existing permutation invariant methods on synthetic and real world datasets.

[What Did You Think Would Happen? Explaining Agent Behaviour through Intended Outcomes](#)

- Herman Yau, Chris Russell, Simon Hadfield
- abstract@[open-review](#): We present a novel form of explanation for Reinforcement Learning, based around the notion of intended outcome. These explanations describe the outcome an agent is trying to achieve by its actions. We provide a simple proof that general methods for post-hoc explanations of this nature are impossible in traditional reinforcement learning. Rather, the information needed for the explanations must be collected in conjunction with training the agent. We derive approaches designed to extract local explanations based on intention for several variants of Q-function approximation and prove consistency between the explanations and the Q-values learned. We demonstrate our method on multiple reinforcement learning problems, and provide code to help researchers introspecting their RL environments and algorithms.

[Batch normalization provably avoids ranks collapse for randomly initialised deep networks](#)

- Hadi Daneshmand, Jonas Kohler, Francis Bach, Thomas Hofmann, Aurelien Lucchi
- abstract@[open-review](#): Randomly initialized neural networks are known to become harder to train with increasing depth, unless architectural enhancements like residual connections and batch normalization are used. We here investigate this phenomenon by revisiting the connection between random initialization in deep networks and spectral instabilities in products of random matrices. Given the rich literature on random matrices, it is not surprising to find that the rank of the intermediate representations in unnormalized networks collapses quickly with depth. In this work we highlight the fact that batch normalization is an effective strategy to avoid rank collapse for both linear and ReLU networks. Leveraging tools from Markov chain theory, we derive a meaningful lower rank bound in deep linear networks. Empirically, we also demonstrate that this rank robustness generalizes to ReLU nets. Finally, we conduct an extensive set of experiments on real-world data sets, which confirm that rank stability is indeed a crucial condition for training modern-day deep neural architectures.

[Choice Bandits](#)

- Arpit Agarwal, Nicholas Johnson, Shivani Agarwal
- abstract@[open-review](#): There has been much interest in recent years in the problem of dueling bandits, where on each round the learner plays a pair of arms and receives as feedback the outcome of a relative pairwise comparison between them. Here we study a natural generalization, that we term \emph{choice bandits}, where the learner plays a set of up to $\$k \geq 2\$$ arms and receives limited relative feedback in the form of a single multiway choice among the pulled arms, drawn from an underlying multiway choice model. We study choice bandits under a very general class of choice models that is characterized by the existence of a unique best' arm (which we term generalized Condorcet winner), and includes as special cases the well-studied multinomial logit (MNL) and multinomial probit (MNP) choice models, and more generally, the class of random utility models with i.i.d. noise (IID-RUMs). We propose an algorithm for choice bandits, termed Winner Beats All (WBA), with distribution dependent $\$O(\log T)\$$ regret bound under all these choice models. The challenge in our setting is that the decision space is $\$|\Theta(n^k)|\$$, which is large for even moderate $\$k\$$. Our algorithm addresses this challenge by extracting just $\$O(n^2)\$$ statistics from multiway choices and exploiting the existence of a uniquebest' arm to find arms that are competitive to this arm in order to construct sets with low regret. Since these statistics are extracted from the same choice observations, one needs a careful martingale analysis in order to show that these statistics are concentrated. We complement our upper bound result with a lower bound result, which shows that our upper bound is order-wise optimal. Our experiments demonstrate that for the special case of $\$k=2\$$, our algorithm is competitive when compared to previous dueling bandit algorithms, and for the more general case $\$k>2\$$, outperforms the recently proposed MaxMinUCB algorithm designed for the MNL model.

[What if Neural Networks had SVDs?](#)

- Alexander Mathiasen, Frederik Vilshøj, Jakob Rødsaaard Jørgensen, Anshul Nasery, Davide Mottin
- abstract@[open-review](#): Various Neural Networks employ time-consuming matrix operations like matrix inversion. Many such matrix operations are faster to compute given the Singular Value Decomposition (SVD). Techniques from (Zhang et al., 2018; Mhammedi et al., 2017) allow using the SVD in Neural Networks without computing it. In theory, the techniques can speed up matrix operations, however, in practice, they are not fast enough. We present an algorithm that is fast enough to speed up several matrix operations. The algorithm increases the degree of parallelism of an underlying matrix multiplication H^*X where H is an orthogonal matrix represented by a product of Householder matrices.

[A Matrix Chernoff Bound for Markov Chains and Its Application to Co-occurrence Matrices](#)

- Jiezhong Qiu, Chi Wang, Ben Liao, Richard Peng, Jie Tang
- abstract@[open-review](#): We prove a Chernoff-type bound for sums of matrix-valued random variables sampled via a regular (aperiodic and irreducible) finite Markov chain. Specially, consider a random walk on a regular Markov chain and a Hermitian matrix-valued function on its state space. Our result gives exponentially decreasing bounds on the tail distributions of the extreme eigenvalues of the sample mean matrix. Our proof is based on the matrix expander (regular undirected graph) Chernoff bound [Garg et al. STOC '18] and scalar Chernoff-Hoeffding bounds for Markov chains [Chung et al. STACS '12]. Our matrix Chernoff bound for Markov chains can be applied to analyze the behavior of co-occurrence statistics for sequential data, which have been common and important data signals in machine learning. We show that given a regular Markov chain with n states and mixing time t , we need a trajectory of length $O(t(\log(n) + \log(t))/\epsilon^2)$ to achieve an estimator of the co-occurrence matrix with error bound ϵ . We conduct several experiments and the experimental results are consistent with the exponentially fast convergence rate from theoretical analysis. Our result gives the first bound on the convergence rate of the co-occurrence matrix and the first sample complexity analysis in graph representation learning.

[CoMIR: Contrastive Multimodal Image Representation for Registration](#)

- Nicolas Pielawski, Elisabeth Wetzer, Johan Åverstedt, Jiahao Lu, Carolina Wählby, Joakim Lindblad, Natasa Sladoje
- abstract@[open-review](#): We propose contrastive coding to learn shared, dense image representations, referred to as CoMIRs (Contrastive Multimodal Image Representations). CoMIRs enable the registration of multimodal images where existing registration methods often fail due to a lack of sufficiently similar image structures. CoMIRs reduce the multimodal registration problem to a monomodal one, in which general intensity-based, as well as feature-based, registration algorithms can be applied. The method involves training one neural network per modality on aligned images, using a contrastive loss based on noise-contrastive estimation (InfoNCE). Unlike other contrastive coding methods, used for, e.g., classification, our approach generates image-like representations that contain the information shared between modalities. We introduce a novel, hyperparameter-free modification to InfoNCE, to enforce rotational equivariance of the learnt representations, a property essential to the registration task. We assess the extent of achieved rotational

equivariance and the stability of the representations with respect to weight initialization, training set, and hyperparameter settings, on a remote sensing dataset of RGB and near-infrared images. We evaluate the learnt representations through registration of a biomedical dataset of bright-field and second-harmonic generation microscopy images; two modalities with very little apparent correlation. The proposed approach based on CoMIRs significantly outperforms registration of representations created by GAN-based image-to-image translation, as well as a state-of-the-art, application-specific method which takes additional knowledge about the data into account. Code is available at: <https://github.com/MIDA-group/CoMIR>.

[Ensuring Fairness Beyond the Training Data](#)

- Debmalya Mandal, Samuel Deng, Suman Jana, Jeannette Wing, Daniel J. Hsu
- abstract@[open-review](#): We initiate the study of fair classifiers that are robust to perturbations in the training distribution. Despite recent progress, the literature on fairness has largely ignored the design of fair and robust classifiers. In this work, we develop classifiers that are fair not only with respect to the training distribution but also for a class of distributions that are weighted perturbations of the training samples. We formulate a min-max objective function whose goal is to minimize a distributionally robust training loss, and at the same time, find a classifier that is fair with respect to a class of distributions. We first reduce this problem to finding a fair classifier that is robust with respect to the class of distributions. Based on an online learning algorithm, we develop an iterative algorithm that provably converges to such a fair and robust solution. Experiments on standard machine learning fairness datasets suggest that, compared to the state-of-the-art fair classifiers, our classifier retains fairness guarantees and test accuracy for a large class of perturbations on the test set. Furthermore, our experiments show that there is an inherent trade-off between fairness robustness and accuracy of such classifiers.

[How do fair decisions fare in long-term qualification?](#)

- Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, Cheng Zhang
- abstract@[open-review](#): Although many fairness criteria have been proposed for decision making, their long-term impact on the well-being of a population remains unclear. In this work, we study the dynamics of population qualification and algorithmic decisions under a partially observed Markov decision problem setting. By characterizing the equilibrium of such dynamics, we analyze the long-term impact of static fairness constraints on the equality and improvement of group well-being. Our results show that static fairness constraints can either promote equality or exacerbate disparity depending on the driving factor of qualification transitions and the effect of sensitive attributes on feature distributions. We also consider possible interventions that can effectively improve group qualification or promote equality of group qualification. Our theoretical results and experiments on static real-world datasets with simulated dynamics show that our framework can be used to facilitate social science studies.

[Pre-training via Paraphrasing](#)

- Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, Luke Zettlemoyer
- abstract@[open-review](#): We introduce MARGE, a pre-trained sequence-to-sequence model learned with an unsupervised multi-lingual multi-document paraphrasing objective. MARGE provides an alternative to the dominant masked language modeling paradigm, where we self-supervise the \emph{reconstruction} of target text by \emph{retrieving} a set of related texts (in many languages) and conditioning on them to maximize the likelihood of generating the original. We show it is possible to jointly learn to do retrieval and reconstruction, given only a random initialization. The objective noisily captures aspects of paraphrase, translation, multi-document summarization, and information retrieval, allowing for strong zero-shot performance on several tasks. For example, with no additional task-specific training we achieve BLEU scores of up to 35.8 for document translation. We further show that fine-tuning gives strong performance on a range of discriminative and generative tasks in many languages, making MARGE the most generally applicable pre-training method to date.

[GCN meets GPU: Decoupling â€œWhen to Sampleâ€ from â€œHow to Sampleâ€](#)

- Morteza Ramezani, Weilin Cong, Mehrdad Mahdavi, Anand Sivasubramaniam, Mahmut Kandemir
- abstract@[open-review](#): Sampling-based methods promise scalability improvements when paired with stochastic gradient descent in training Graph Convolutional Networks (GCNs). While effective in alleviating the neighborhood explosion, due to bandwidth and memory bottlenecks, these methods lead to computational overheads in preprocessing and loading new samples in heterogeneous systems, which significantly deteriorate the sampling performance. By decoupling the frequency of sampling from the sampling strategy, we propose LazyGCN, a general yet effective framework that can be integrated with any sampling strategy to substantially improve the training time. The basic idea behind LazyGCN is to perform sampling periodically and effectively recycle the sampled nodes to mitigate data preparation overhead. We theoretically analyze the proposed algorithm and show that under a mild condition on the recycling size, by reducing the variance of inner layers, we are able to obtain the same convergence rate as the underlying sampling method. We also give corroborating empirical evidence on large real-world graphs, demonstrating that the proposed schema can significantly reduce the number of sampling steps and yield superior speedup without compromising the accuracy.

[Continual Learning of a Mixed Sequence of Similar and Dissimilar Tasks](#)

- Zixuan Ke, Bing Liu, Xingchang Huang
- abstract@[open-review](#): Existing research on continual learning of a sequence of tasks focused on dealing with catastrophic forgetting, where the tasks are assumed to be dissimilar and have little shared knowledge. Some work has also been done to transfer previously learned knowledge to the new task when the tasks are similar and have shared knowledge. However, in the most general case, a CL system not only should have the above two capabilities, but also the \textit{backward knowledge transfer} capability so that future tasks may help improve the past models whenever possible. To the best of our knowledge, no technique has been proposed to learn a sequence of mixed similar and dissimilar tasks that can deal with forgetting and also transfer knowledge forward and backward. This paper proposes such a technique to learn both types of tasks in the same network. For dissimilar tasks, the algorithm focuses on dealing with forgetting, and for similar tasks, the algorithm focuses on selectively transferring the knowledge learned from some similar previous tasks to improve the new task learning. Additionally, the algorithm automatically detects whether a new task is similar to any previous tasks. Empirical evaluation using sequences of mixed tasks demonstrates the effectiveness of the proposed model.

[All your loss are belong to Bayes](#)

- Christian Walder, Richard Nock
- abstract@[open-review](#): Loss functions are a cornerstone of machine learning and the starting point of most algorithms. Statistics and Bayesian decision theory have contributed, via properness, to elicit over the past decades a wide set of admissible losses in supervised learning, to which most popular choices belong (logistic, square, Matsushita, etc.). Rather than making a potentially biased ad hoc choice of the loss, there has recently been a boost in efforts to fit the loss to the domain at hand while training the model itself. The key approaches fit a canonical link, a function which monotonically relates the closed unit interval to \mathbb{R} and can provide a proper loss via integration. In this paper, we rely on a broader view of proper composite losses and a recent construct from information geometry, source functions, whose fitting alleviates constraints faced by canonical links. We introduce a trick on squared Gaussian Processes to obtain a random process whose paths are compliant source functions with many desirable properties in the context of link estimation. Experimental results demonstrate substantial improvements over the state of the art.

[HAWQ-V2: Hessian Aware trace-Weighted Quantization of Neural Networks](#)

- Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, Kurt Keutzer
- abstract@[open-review](#): Quantization is an effective method for reducing memory footprint and inference time of Neural Networks. However, ultra low precision quantization could lead to significant degradation in model accuracy. A promising method to address this is to perform mixed-precision quantization, where more sensitive layers are kept at higher precision. However, the search space for a mixed-precision quantization is exponential in the number of layers. Recent work has proposed a novel Hessian based framework, with the aim of reducing this exponential search space by using second-order information. While promising, this prior work has three major limitations: (i) they only use a heuristic metric based on top Hessian eigenvalue as a measure of sensitivity and do not consider the rest of the Hessian spectrum; (ii) their approach only provides relative sensitivity of different layers and therefore requires a manual selection of the mixed-precision setting; and (iii) they do not consider mixed-precision activation quantization. Here, we present HAWQ-V2 which addresses these shortcomings. For (i), we theoretically prove that the right sensitivity metric is the average Hessian trace, instead of just top Hessian eigenvalue. For (ii), we develop a Pareto frontier based method for automatic bit precision selection of different layers without any manual intervention. For (iii), we develop the first Hessian based analysis for mixed-precision activation quantization, which is very beneficial for object detection. We show that HAWQ-V2 achieves new state-of-the-art results for a wide range of tasks. In particular, we present quantization results for InceptionV3, ResNet50, and SqueezeNext, all without any manual bit selection. Furthermore, we present results for object detection on Microsoft COCO, where we achieve 2.6 higher mAP than direct uniform quantization and 1.6 higher mAP than the recently proposed method of FQN, with a smaller model size of 17.9MB.

[Sample-Efficient Reinforcement Learning of Undercomplete POMDPs](#)

- Chi Jin, Sham Kakade, Akshay Krishnamurthy, Qinghua Liu
- abstract@[open-review](#): Partial observability is a common challenge in many reinforcement learning applications, which requires an agent to maintain memory, infer latent states, and integrate this past information into exploration. This challenge leads to a number of computational and statistical hardness results for learning general Partially Observable Markov Decision Processes (POMDPs). This work shows that these hardness barriers do not preclude efficient reinforcement learning for rich and interesting subclasses of POMDPs. In particular, we present a sample-efficient algorithm, OOM-UCB, for episodic finite undercomplete POMDPs, where the number of observations is larger than the number of latent states and where exploration is essential for learning, thus distinguishing our results from prior works. OOM-UCB achieves an optimal sample complexity of $\tilde{O}(1/\epsilon^2)$ for finding an ϵ -optimal policy, along with being polynomial in all other relevant quantities. As an interesting special case, we also provide a computationally and statistically efficient algorithm for POMDPs with deterministic state transitions.

[Non-Convex SGD Learns Halfspaces with Adversarial Label Noise](#)

- Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, Nikos Zarifis
- abstract@[open-review](#): We study the problem of agnostically learning homogeneous halfspaces in the distribution-specific PAC model. For a broad family of structured distributions, including log-concave distributions, we show that non-convex SGD efficiently converges to a solution with misclassification error $O(\text{opt}) + \epsilon$, where opt is the misclassification error of the best-fitting halfspace. In sharp contrast, we show that optimizing any convex surrogate inherently leads to misclassification error of $\Omega(\text{opt})$, even under Gaussian marginals.

[A Tight Lower Bound and Efficient Reduction for Swap Regret](#)

- Shinji Ito
- abstract@[open-review](#): Swap regret, a generic performance measure of online decision-making algorithms, plays an important role in the theory of repeated games, along with a close connection to correlated equilibria in strategic games. This paper shows an $\Omega(\sqrt{T N \log N})$ -lower bound for swap regret, where T and N denote the numbers of time steps and available actions, respectively. Our lower bound is tight up to a constant, and resolves an open problem mentioned, e.g., in the book by Nisan et al. (2007). Besides, we present a computationally efficient reduction method that converts no-external-regret algorithms to no-swap-regret algorithms. This method can be applied not only to the full-information setting but also to the bandit setting and provides a better regret bound than previous results.

[DisCor: Corrective Feedback in Reinforcement Learning via Distribution Correction](#)

- Aviral Kumar, Abhishek Gupta, Sergey Levine
- abstract@[open-review](#): Deep reinforcement learning can learn effective policies for a wide range of tasks, but is notoriously difficult to use due to instability and sensitivity to hyperparameters. The reasons for this remain unclear. In this paper, we study how RL methods based on bootstrapping-based Q-learning can suffer from a pathological interaction between function approximation and the data distribution used to train the Q-function: with standard supervised learning, online data collection should induce corrective feedback, where new data corrects mistakes in old predictions. With dynamic programming methods like Q-learning, such feedback may be absent. This can lead to potential instability, sub-optimal convergence, and poor results when learning from noisy, sparse or delayed rewards. Based on these observations, we propose a new algorithm, DisCor, which explicitly optimizes for data distributions that can correct for accumulated errors in the value function. DisCor computes a tractable approximation to the distribution that optimally induces corrective feedback, which we show results in reweighting samples based on the estimated accuracy of their target values. Using this distribution for training, DisCor results in substantial improvements in a range of challenging RL settings, such as multi-task learning and learning from noisy reward signals.

[OTLDA: A Geometry-aware Optimal Transport Approach for Topic Modeling](#)

- Viet Huynh, He Zhao, Dinh Phung
- abstract@[open-review](#): We present an optimal transport framework for learning topics from textual data. While the celebrated Latent Dirichlet allocation (LDA) topic model and its variants have been applied to many disciplines, they mainly focus on word-occurrences and neglect to incorporate semantic regularities in language. Even though recent works have tried to exploit the semantic relationship between words to bridge this gap, however, these models which are usually extensions of LDA or Dirichlet Multinomial mixture (DMM) are tailored to deal effectively with either regular or short documents. The optimal transport distance provides an appealing tool to incorporate the geometry of word semantics into it. Moreover, recent developments on efficient computation of optimal transport distance also promote its application in topic modeling. In this paper we ground on optimal transport theory to naturally exploit the geometric structures of semantically related words in embedding spaces which leads to more interpretable learned topics. Comprehensive experiments illustrate that the proposed framework outperforms competitive approaches in terms of topic coherence on assorted text corpora which include both long and short documents. The representation of learned topic also leads to better accuracy on classification downstream tasks, which is considered as an extrinsic evaluation.

[Measuring Robustness to Natural Distribution Shifts in Image Classification](#)

- Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, Ludwig Schmidt
- abstract@[open-review](#): We study how robust current ImageNet models are to distribution shifts arising from natural variations in datasets. Most research on robustness focuses on synthetic image perturbations (noise, simulated weather artifacts, adversarial examples, etc.), which leaves open how robustness on synthetic distribution shift relates to distribution shift arising in real data. Informed by an evaluation of 204 ImageNet models in 213 different test conditions, we find that there is often little to no transfer of robustness from current synthetic to natural distribution shift. Moreover, most current techniques provide no robustness to the natural distribution shifts in our testbed. The main exception is training on larger and more diverse datasets, which

in multiple cases increases robustness, but is still far from closing the performance gaps. Our results indicate that distribution shifts arising in real data are currently an open research problem.

[Can I Trust My Fairness Metric? Assessing Fairness with Unlabeled Data and Bayesian Inference](#)

- Disi Ji, Padhraic Smyth, Mark Steyvers
- abstract@[open-review](#): Group fairness is measured via parity of quantitative metrics across different protected demographic groups. In this paper, we investigate the problem of reliably assessing group fairness metrics when labeled examples are few but unlabeled examples are plentiful. We propose a general Bayesian framework that can augment labeled data with unlabeled data to produce more accurate and lower-variance estimates compared to methods based on labeled data alone. Our approach estimates calibrated scores (for unlabeled examples) of each group using a hierarchical latent variable model conditioned on labeled examples. This in turn allows for inference of posterior distributions for an array of group fairness metrics with a notion of uncertainty. We demonstrate that our approach leads to significant and consistent reductions in estimation error across multiple well-known fairness datasets, sensitive attributes, and predictive models. The results clearly show the benefits of using both unlabeled data and Bayesian inference in assessing whether a prediction model is fair or not.

[RandAugment: Practical Automated Data Augmentation with a Reduced Search Space](#)

- Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, Quoc Le
- abstract@[open-review](#): Recent work on automated data augmentation strategies has led to state-of-the-art results in image classification and object detection. An obstacle to a large-scale adoption of these methods is that they require a separate and expensive search phase. A common way to overcome the expense of the search phase was to use a smaller proxy task. However, it was not clear if the optimized hyperparameters found on the proxy task are also optimal for the actual task. In this work, we rethink the process of designing automated data augmentation strategies. We find that while previous work required searching for many augmentation parameters (e.g. magnitude and probability) independently for each augmentation operation, it is sufficient to only search for a single parameter that jointly controls all operations. Hence, we propose a search space that is vastly smaller (e.g. from 10^{32} to 10^2 potential candidates). The smaller search space significantly reduces the computational expense of automated data augmentation and permits the removal of a separate proxy task. Despite the simplifications, our method achieves state-of-the-art performance on CIFAR-10, SVHN, and ImageNet. On EfficientNet-B7, we achieve 84.7% accuracy, a 1.0% increase over baseline augmentation and a 0.4% improvement over AutoAugment on the ImageNet dataset. On object detection, the same method used for classification leads to 1.0-1.3% improvement over the baseline augmentation method on COCO. Code is available online.

[Asymptotic normality and confidence intervals for derivatives of 2-layers neural network in the random features model](#)

- Yiwei Shen, Pierre C Belloc
- abstract@[open-review](#): This paper studies two-layers Neural Networks (NN), where the first layer contains random weights, and the second layer is trained using Ridge regularization. This model has been the focus of numerous recent works, showing that despite its simplicity, it captures some of the empirically observed behaviors of NN in the overparametrized regime, such as the double-descent curve where the generalization error decreases as the number of weights increases to ∞ . This paper establishes asymptotic distribution results for this 2-layers NN model in the regime where the ratios $\frac{p}{n}$ and $\frac{d}{n}$ have finite limits, where n is the sample size, p the ambient dimension and d is the width of the first layer. We show that a weighted average of the derivatives of the trained NN at the observed data is asymptotically normal, in a setting with Lipschitz activation functions in a linear regression response with Gaussian features under possibly non-linear perturbations. We then leverage this asymptotic normality result to construct confidence intervals (CIs) for single components of the unknown regression vector. The novelty of our results are threefold: (1) Despite the nonlinearity induced by the activation function, we characterize the asymptotic distribution of a weighted average of the gradients of the network after training; (2) It provides the first frequentist uncertainty quantification guarantees, in the form of valid $(1-\alpha)$ -CIs, based on NN estimates; (3) It shows that the double-descent phenomenon occurs in terms of the length of the CIs, with the length increasing and then decreasing as $\frac{d}{n} \rightarrow \infty$ for certain fixed values of $\frac{p}{n}$. We also provide a toolbox to predict the length of CIs numerically, which lets us compare activation functions and other parameters in terms of CI length.

[DisARM: An Antithetic Gradient Estimator for Binary Latent Variables](#)

- Zhe Dong, Andriy Mnih, George Tucker
- abstract@[open-review](#): Training models with discrete latent variables is challenging due to the difficulty of estimating the gradients accurately. Much of the recent progress has been achieved by taking advantage of continuous relaxations of the system, which are not always available or even possible. The Augment-REINFORCE-Merge (ARM) estimator provides an alternative that, instead of relaxation, uses continuous augmentation. Applying antithetic sampling over the augmenting variables yields a relatively low-variance and unbiased estimator applicable to any model with binary latent variables. However, while antithetic sampling reduces variance, the augmentation process increases variance. We show that ARM can be improved by analytically integrating out the randomness introduced by the augmentation process, guaranteeing substantial variance reduction. Our estimator, DisARM, is simple to implement and has the same computational cost as ARM. We evaluate DisARM on several generative modeling benchmarks and show that it consistently outperforms ARM and a strong independent sample baseline in terms of both variance and log-likelihood. Furthermore, we propose a local version of DisARM designed for optimizing the multi-sample variational bound, and show that it outperforms VIMCO, the current state-of-the-art method.

[Variational Inference for Graph Convolutional Networks in the Absence of Graph Data and Adversarial Settings](#)

- Pantelis Elinas, Edwin V. Bonilla, Louis Tiao
- abstract@[open-review](#): We propose a framework that lifts the capabilities of graph convolutional networks (GCNs) to scenarios where no input graph is given and increases their robustness to adversarial attacks. We formulate a joint probabilistic model that considers a prior distribution over graphs along with a GCN-based likelihood and develop a stochastic variational inference algorithm to estimate the graph posterior and the GCN parameters jointly. To address the problem of propagating gradients through latent variables drawn from discrete distributions, we use their continuous relaxations known as Concrete distributions. We show that, on real datasets, our approach can outperform state-of-the-art Bayesian and non-Bayesian graph neural network algorithms on the task of semi-supervised classification in the absence of graph data and when the network structure is subjected to adversarial perturbations.

[Supervised Contrastive Learning](#)

- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, Dilip Krishnan
- abstract@[open-review](#): Contrastive learning applied to self-supervised representation learning has seen a resurgence in recent years, leading to state of the art performance in the unsupervised training of deep image models. Modern batch contrastive approaches subsume or significantly outperform traditional contrastive losses such as triplet, max-margin and the N-pairs loss. In this work, we extend the self-supervised batch contrastive approach to the fully-supervised setting, allowing us to effectively leverage label information. Clusters of points belonging to the same class are pulled together in embedding space, while simultaneously pushing apart clusters of samples from different classes. We analyze two possible versions of the supervised contrastive (SupCon) loss, identifying the best-performing formulation of the loss. On ResNet-200, we achieve top-1 accuracy of 81.4% on the ImageNet dataset, which is 0.8% above the best number reported for this architecture. We show consistent outperformance over cross-entropy on other datasets and two ResNet variants. The loss shows benefits for robustness to natural corruptions, and is more stable to hyperparameter settings such as optimizers and data

augmentations. In reduced data settings, it outperforms cross-entropy significantly. Our loss function is simple to implement and reference TensorFlow code is released at <https://t.ly/supcon>.

[Learning Optimal Representations with the Decodable Information Bottleneck](#)

- Yann Dubois, Douwe Kiela, David J. Schwab, Ramakrishna Vedantam
- abstract@[open-review](#): We address the question of characterizing and finding optimal representations for supervised learning. Traditionally, this question has been tackled using the Information Bottleneck, which compresses the inputs while retaining information about the targets, in a decoder-agnostic fashion. In machine learning, however, our goal is not compression but rather generalization, which is intimately linked to the predictive family or decoder of interest (e.g. linear classifier). We propose the Decodable Information Bottleneck (DIB) that considers information retention and compression from the perspective of the desired predictive family. As a result, DIB gives rise to representations that are optimal in terms of expected test performance and can be estimated with guarantees. Empirically, we show that the framework can be used to enforce a small generalization gap on downstream classifiers and to predict the generalization ability of neural networks.

[Meta-trained agents implement Bayes-optimal agents](#)

- Vladimir Mikulik, GrÃ©goire DelÃ©tang, Tom McGrath, Tim Genewein, Miljan Martic, Shane Legg, Pedro Ortega
- abstract@[open-review](#): Memory-based meta-learning is a powerful technique to build agents that adapt fast to any task within a target distribution. A previous theoretical study has argued that this remarkable performance is because the meta-training protocol incentivises agents to behave Bayes-optimally. We empirically investigate this claim on a number of prediction and bandit tasks. Inspired by ideas from theoretical computer science, we show that meta-learned and Bayes-optimal agents not only behave alike, but they even share a similar computational structure, in the sense that one agent system can approximately simulate the other. Furthermore, we show that Bayes-optimal agents are fixed points of the meta-learning dynamics. Our results suggest that memory-based meta-learning is a general technique for numerically approximating Bayes-optimal agents; that is, even for task distributions for which we currently don't possess tractable models.

[Learning Agent Representations for Ice Hockey](#)

- Guiliang Liu, Oliver Schulte, Pascal Poupart, Mike Rudd, Mehrsan Javan
- abstract@[open-review](#): Team sports is a new application domain for agent modeling with high real-world impact. A fundamental challenge for modeling professional players is their large number (over 1K), which includes many bench players with sparse participation in a game season. The diversity and sparsity of player observations make it difficult to extend previous agent representation models to the sports domain. This paper develops a new approach for agent representations, based on a Markov game model, that is tailored towards applications in professional ice hockey. We introduce a novel player representation via player generation framework where a variational encoder embeds player information with latent variables. The encoder learns a context-specific shared prior to induce a shrinkage effect for the posterior player representations, allowing it to share statistical information across players with different participations. To model the play dynamics in sequential sports data, we design a Variational Recurrent Ladder Agent Encoder (VaRLAE). It learns a contextualized player representation with a hierarchy of latent variables that effectively prevents latent posterior collapse. We validate our player representations in major sports analytics tasks. Our experimental results, based on a large dataset that contains over 4.5M events, show state-of-the-art performance for our VarLAE on facilitating 1) identifying the acting player, 2) estimating expected goals, and 3) predicting the final score difference.

[Weak Form Generalized Hamiltonian Learning](#)

- Kevin Course, Trefor Evans, Prasanth Nair
- abstract@[open-review](#): We present a method for learning generalized Hamiltonian decompositions of ordinary differential equations given a set of noisy time series measurements. Our method simultaneously learns a continuous time model and a scalar energy function for a general dynamical system. Learning predictive models in this form allows one to place strong, high-level, physics inspired priors onto the form of the learnt governing equations for general dynamical systems. Moreover, having shown how our method extends and unifies some previous work in deep learning with physics inspired priors, we present a novel method for learning continuous time models from the weak form of the governing equations which is less computationally taxing than standard adjoint methods.

[Neural Non-Rigid Tracking](#)

- Aljaz Bozic, Pablo Palafox, Michael ZollhÃ¶fer, Angela Dai, Justus Thies, Matthias Niessner
- abstract@[open-review](#): We introduce a novel, end-to-end learnable, differentiable non-rigid tracker that enables state-of-the-art non-rigid reconstruction by a learned robust optimization. Given two input RGB-D frames of a non-rigidly moving object, we employ a convolutional neural network to predict dense correspondences and their confidences. These correspondences are used as constraints in an as-rigid-as-possible (ARAP) optimization problem. By enabling gradient back-propagation through the weighted non-linear least squares solver, we are able to learn correspondences and confidences in an end-to-end manner such that they are optimal for the task of non-rigid tracking. Under this formulation, correspondence confidences can be learned via self-supervision, informing a learned robust optimization, where outliers and wrong correspondences are automatically down-weighted to enable effective tracking. Compared to state-of-the-art approaches, our algorithm shows improved reconstruction performance, while simultaneously achieving 85 times faster correspondence prediction than comparable deep-learning based methods.

[Collegial Ensembles](#)

- Etai Littwin, Ben Myara, Sima Sabah, Joshua Susskind, Shuangfei Zhai, Oren Golan
- abstract@[open-review](#): Modern neural network performance typically improves as model size increases. A recent line of research on the Neural Tangent Kernel (NTK) of over-parameterized networks indicates that the improvement with size increase is a product of a better conditioned loss landscape. In this work, we investigate a form of over-parameterization achieved through ensembling, where we define collegial ensembles (CE) as the aggregation of multiple independent models with identical architectures, trained as a single model. We show that the optimization dynamics of CE simplify dramatically when the number of models in the ensemble is large, resembling the dynamics of wide models, yet scale much more favorably. We use recent theoretical results on the finite width corrections of the NTK to perform efficient architecture search in a space of finite width CE that aims to either minimize capacity, or maximize trainability under a set of constraints. The resulting ensembles can be efficiently implemented in practical architectures using group convolutions and block diagonal layers. Finally, we show how our framework can be used to analytically derive optimal group convolution modules originally found using expensive grid searches, without having to train a single model.

[ICNet: Intra-saliency Correlation Network for Co-Saliency Detection](#)

- Wen-Da Jin, Jun Xu, Ming-Ming Cheng, Yi Zhang, Wei Guo
- abstract@[open-review](#): Intra-saliency and inter-saliency cues have been extensively studied for co-saliency detection (Co-SOD). Model-based methods produce coarse Co-SOD results due to hand-crafted intra- and inter-saliency features. Current data-driven models exploit inter-saliency cues, but undervalue the potential power of intra-saliency cues. In this paper, we propose an Intra-saliency Correlation Network (ICNet) to extract intra-saliency cues from the single image saliency maps (SISMs) predicted by any off-the-shelf SOD method, and obtain inter-saliency cues by correlation techniques.

Specifically, we adopt normalized masked average pooling (NMAP) to extract latent intra-saliency categories from the SISMs and semantic features as intra cues. Then we employ a correlation fusion module (CFM) to obtain inter cues by exploiting correlations between the intra cues and single-image features. To improve Co-SOD performance, we propose a category-independent rearranged self-correlation feature (RSCF) strategy. Experiments on three benchmarks show that our ICNet outperforms previous state-of-the-art methods on Co-SOD. Ablation studies validate the effectiveness of our contributions. The PyTorch code is available at <https://github.com/blanclist/ICNet>.

[Improved Variational Bayesian Phylogenetic Inference with Normalizing Flows](#)

- Cheng Zhang
- abstract@[open-review](#): Variational Bayesian phylogenetic inference (VBPI) provides a promising general variational framework for efficient estimation of phylogenetic posteriors. However, the current diagonal Lognormal branch length approximation would significantly restrict the quality of the approximating distributions. In this paper, we propose a new type of VBPI, VBPI-NF, as a first step to empower phylogenetic posterior estimation with deep learning techniques. By handling the non-Euclidean branch length space of phylogenetic models with carefully designed permutation equivariant transformations, VBPI-NF uses normalizing flows to provide a rich family of flexible branch length distributions that generalize across different tree topologies. We show that VBPI-NF significantly improves upon the vanilla VBPI on a benchmark of challenging real data Bayesian phylogenetic inference problems. Further investigation also reveals that the structured parameterization in those permutation equivariant transformations can provide additional amortization benefit.

[Deep Metric Learning with Spherical Embedding](#)

- Dingyi Zhang, Yingming Li, Zhongfei Zhang
- abstract@[open-review](#): Deep metric learning has attracted much attention in recent years, due to seamlessly combining the distance metric learning and deep neural network. Many endeavors are devoted to design different pair-based angular loss functions, which decouple the magnitude and direction information for embedding vectors and ensure the training and testing measure consistency. However, these traditional angular losses cannot guarantee that all the sample embeddings are on the surface of the same hypersphere during the training stage, which would result in unstable gradient in batch optimization and may influence the quick convergence of the embedding learning. In this paper, we first investigate the effect of the embedding norm for deep metric learning with angular distance, and then propose a spherical embedding constraint (SEC) to regularize the distribution of the norms. SEC adaptively adjusts the embeddings to fall on the same hypersphere and performs more balanced direction update. Extensive experiments on deep metric learning, face recognition, and contrastive self-supervised learning show that the SEC-based angular space learning strategy significantly improves the performance of the state-of-the-art.

[Preference-based Reinforcement Learning with Finite-Time Guarantees](#)

- Yichong Xu, Ruosong Wang, Lin Yang, Aarti Singh, Artur Dubrawski
- abstract@[open-review](#): Preference-based Reinforcement Learning (PbRL) replaces reward values in traditional reinforcement learning by preferences to better elicit human opinion on the target objective, especially when numerical reward values are hard to design or interpret. Despite promising results in applications, the theoretical understanding of PbRL is still in its infancy. In this paper, we present the first finite-time analysis for general PbRL problems. We first show that a unique optimal policy may not exist if preferences over trajectories are deterministic for PbRL. If preferences are stochastic, and the preference probability relates to the hidden reward values, we present algorithms for PbRL, both with and without a simulator, that are able to identify the best policy up to accuracy $\$w\backslash epsilon$ with high probability. Our method explores the state space by navigating to under-explored states, and solves PbRL using a combination of dueling bandits and policy search. Experiments show the efficacy of our method when it is applied to real-world problems.

[AdaBelief Optimizer: Adapting Stepsizes by the Belief in Observed Gradients](#)

- Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar C. Tatikonda, Nicha Dvornek, Xenophon Papademetris, James Duncan
- abstract@[open-review](#): Most popular optimizers for deep learning can be broadly categorized as adaptive methods (e.g., Adam) and accelerated schemes (e.g., stochastic gradient descent (SGD) with momentum). For many models such as convolutional neural networks (CNNs), adaptive methods typically converge faster but generalize worse compared to SGD; for complex settings such as generative adversarial networks (GANs), adaptive methods are typically the default because of their stability. We propose AdaBelief to simultaneously achieve three goals: fast convergence as in adaptive methods, good generalization as in SGD, and training stability. The intuition for AdaBelief is to adapt the stepsize according to the "belief" in the current gradient direction. Viewing the exponential moving average (EMA) of the noisy gradient as the prediction of the gradient at the next time step, if the observed gradient greatly deviates from the prediction, we distrust the current observation and take a small step; if the observed gradient is close to the prediction, we trust it and take a large step. We validate AdaBelief in extensive experiments, showing that it outperforms other methods with fast convergence and high accuracy on image classification and language modeling. Specifically, on ImageNet, AdaBelief achieves comparable accuracy to SGD. Furthermore, in the training of a GAN on Cifar10, AdaBelief demonstrates high stability and improves the quality of generated samples compared to a well-tuned Adam optimizer.

[Interpretable Sequence Learning for Covid-19 Forecasting](#)

- Sercan Arik, Chun-Liang Li, Jinsung Yoon, Rajarishi Sinha, Arkady Epshteyn, Long Le, Vikas Menon, Shashank Singh, Leyou Zhang, Martin Nikoltchev, Yash Sonthalia, Hootan Nakhost, Elli Kanal, Tomas Pfister
- abstract@[open-review](#): We propose a novel approach that integrates machine learning into compartmental disease modeling (e.g., SEIR) to predict the progression of COVID-19. Our model is explainable by design as it explicitly shows how different compartments evolve and it uses interpretable encoders to incorporate covariates and improve performance. Explainability is valuable to ensure that the model's forecasts are credible to epidemiologists and to instill confidence in end-users such as policy makers and healthcare institutions. Our model can be applied at different geographic resolutions, and we demonstrate it for states and counties in the United States. We show that our model provides more accurate forecasts compared to the alternatives, and that it provides qualitatively meaningful explanatory insights.

[Off-policy Policy Evaluation For Sequential Decisions Under Unobserved Confounding](#)

- Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, Emma Brunskill
- abstract@[open-review](#): When observed decisions depend only on observed features, off-policy policy evaluation (OPE) methods for sequential decision problems can estimate the performance of evaluation policies before deploying them. However, this assumption is frequently violated due to unobserved confounders, unrecorded variables that impact both the decisions and their outcomes. We assess robustness of OPE methods under unobserved confounding by developing worst-case bounds on the performance of an evaluation policy. When unobserved confounders can affect every decision in an episode, we demonstrate that even small amounts of per-decision confounding can heavily bias OPE methods. Fortunately, in a number of important settings found in healthcare, policy-making, and technology, unobserved confounders may directly affect only one of the many decisions made, and influence future decisions/rewards only through the directly affected decision. Under this less pessimistic model of one-decision confounding, we propose an efficient loss-minimization-based procedure for computing worst-case bounds, and prove its statistical consistency. On simulated healthcare examples---management of sepsis and interventions for autistic children---where this is a reasonable model, we demonstrate that our method invalidates non-robust results and provides meaningful certificates of robustness, allowing reliable selection of policies under unobserved confounding.

[Modern Hopfield Networks and Attention for Immune Repertoire Classification](#)

- Michael Widrich, Bernhard Schäfl, Milena Pavlović, Hubert Ramsauer, Lukas Gruber, Markus Holzleitner, Johannes Brandstetter, Geir Kjetil Sandve, Victor Greiff, Sepp Hochreiter, Gápter Klambauer
- abstract@[open-review](#): A central mechanism in machine learning is to identify, store, and recognize patterns. How to learn, access, and retrieve such patterns is crucial in Hopfield networks and the more recent transformer architectures. We show that the attention mechanism of transformer architectures is actually the update rule of modern Hopfield networks that can store exponentially many patterns. We exploit this high storage capacity of modern Hopfield networks to solve a challenging multiple instance learning (MIL) problem in computational biology: immune repertoire classification. In immune repertoire classification, a vast number of immune receptors are used to predict the immune status of an individual. This constitutes a MIL problem with an unprecedentedly massive number of instances, two orders of magnitude larger than currently considered problems, and with an extremely low witness rate. Accurate and interpretable machine learning methods solving this problem could pave the way towards new vaccines and therapies, which is currently a very relevant research topic intensified by the COVID-19 crisis. In this work, we present our novel method DeepRC that integrates transformer-like attention, or equivalently modern Hopfield networks, into deep learning architectures for massive MIL such as immune repertoire classification. We demonstrate that DeepRC outperforms all other methods with respect to predictive performance on large-scale experiments including simulated and real-world virus infection data and enables the extraction of sequence motifs that are connected to a given disease class. Source code and datasets: <https://github.com/ml-jku/DeepRC>

[One Ring to Rule Them All: Certifiably Robust Geometric Perception with Outliers](#)

- Heng Yang, Luca Carlone
- abstract@[open-review](#): We propose the first general and practical framework to design certifiable algorithms for robust geometric perception in the presence of a large amount of outliers. We investigate the use of a truncated least squares (TLS) cost function, which is known to be robust to outliers, but leads to hard, nonconvex, and nonsmooth optimization problems. Our first contribution is to show that “for a broad class of geometric perception problems” TLS estimation can be reformulated as an optimization over the ring of polynomials and Lasserre’s hierarchy of convex moment relaxations is empirically tight at the minimum relaxation order (i.e., certifiably obtains the global minimum of the nonconvex TLS problem). Our second contribution is to exploit the structural sparsity of the objective and constraint polynomials and leverage basis reduction to significantly reduce the size of the semidefinite program (SDP) resulting from the moment relaxation, without compromising its tightness. Our third contribution is to develop scalable dual optimality certifiers from the lens of sums-of-squares (SOS) relaxation, that can compute the suboptimality gap and possibly certify global optimality of any candidate solution (e.g., returned by fast heuristics such as RANSAC or graduated non-convexity). Our dual certifiers leverage Douglas-Rachford Splitting to solve a convex feasibility SDP. Numerical experiments across different perception problems, including single rotation averaging, shape alignment, 3D point cloud and mesh registration, and high-integrity satellite pose estimation, demonstrate the tightness of our relaxations, the correctness of the certification, and the scalability of the proposed dual certifiers to large problems, beyond the reach of current SDP solvers.

[Task-Robust Model-Agnostic Meta-Learning](#)

- Liam Collins, Aryan Mokhtari, Sanjay Shakkottai
- abstract@[open-review](#): Meta-learning methods have shown an impressive ability to train models that rapidly learn new tasks. However, these methods only aim to perform well in expectation over tasks coming from some particular distribution that is typically equivalent across meta-training and meta-testing, rather than considering worst-case task performance. In this work we introduce the notion of “task-robustness” by reformulating the popular Model-Agnostic Meta-Learning (MAML) objective \cite{finn2017model} such that the goal is to minimize the maximum loss over the observed meta-training tasks. The solution to this novel formulation is task-robust in the sense that it places equal importance on even the most difficult and/or rare tasks. This also means that it performs well over all distributions of the observed tasks, making it robust to shifts in the task distribution between meta-training and meta-testing. We present an algorithm to solve the proposed min-max problem, and show that it converges to an ϵ -accurate point at the optimal rate of $O(1/\epsilon^2)$ in the convex setting and to an (ϵ, δ) -stationary point at the rate of $O(\max\{1/\epsilon^5, 1/\delta^5\})$ in nonconvex settings. We also provide an upper bound on the new task generalization error that captures the advantage of minimizing the worst-case task loss, and demonstrate this advantage in sinusoid regression and image classification experiments.

[R-learning in actor-critic model offers a biologically relevant mechanism for sequential decision-making](#)

- Sergey Shuvaev, Sarah Starosta, Duda Kvitsiani, Adam Kepcs, Alexei Koulakov
- abstract@[open-review](#): In real-world settings, we repeatedly decide whether to pursue better conditions or to keep things unchanged. Examples include time investment, employment, entertainment preferences etc. How do we make such decisions? To address this question, the field of behavioral ecology has developed foraging paradigms “the model settings in which human and non-human subjects decided when to leave depleting food resources. Foraging theory, represented by the marginal value theorem (MVT), provided accurate average-case stay-or-leave rules consistent with behaviors of subjects towards depleting resources. Yet, the algorithms underlying individual choices and ways to learn such algorithms remained unclear. In this work, we build interpretable deep actor-critic models to show that R-learning “a reinforcement learning (RL) approach balancing short-term and long-term rewards” is consistent with the way real-life agents may learn making stay-or-leave decisions. Specifically we show that deep R-learning predicts choice patterns consistent with behavior of mice in foraging tasks; its TD error, the training signal in our model, correlates with dopamine activity of ventral tegmental area (VTA) neurons in the brain. Our theoretical and experimental results show that deep R-learning agents leave depleting reward resources when reward intake rates fall below their exponential averages over past trials. This individual-case decision rule, learned within RL and matching the MVT on average, bridges the gap between these major approaches to sequential decision-making. We further argue that our proposed decision rule, resulting from R-learning and consistent with animals’ behavior, is Bayes optimal in dynamic real-world environments. Overall, our work links available sequential decision-making theories including the MVT, RL, and Bayesian approaches to propose the learning mechanism and an optimal decision rule for sequential stay-or-leave choices in natural environments.

[Revisiting Frank-Wolfe for Polytopes: Strict Complementarity and Sparsity](#)

- Dan Garber
- abstract@[open-review](#): In recent years it was proved that simple modifications of the classical Frank-Wolfe algorithm (aka conditional gradient algorithm) for smooth convex minimization over convex and compact polytopes, converge with linear rate, assuming the objective function has the quadratic growth property. However, the rate of these methods depends explicitly on the dimension of the problem which cannot explain their empirical success for large scale problems. In this paper we first demonstrate that already for very simple problems and even when the optimal solution lies on a low-dimensional face of the polytope, such dependence on the dimension cannot be avoided in worst case. We then revisit the addition of a strict complementarity assumption already considered in Wolfe’s classical book \cite{Wolfe1970}, and prove that under this condition, the Frank-Wolfe method with away-steps and line-search converges linearly with rate that depends explicitly only on the dimension of the optimal face, hence providing a significant improvement in case the optimal solution is sparse. We motivate this strict complementarity condition by proving that it implies sparsity-robustness of optimal solutions to noise.

[Fast Convergence of Langevin Dynamics on Manifold: Geodesics meet Log-Sobolev](#)

- Xiao Wang, Qi Lei, Ioannis Panageas

- abstract@[open-review](#): Sampling is a fundamental and arguably very important task with numerous applications in Machine Learning. One approach to sample from a high dimensional distribution e^{-f} for some function f is the Langevin Algorithm (LA). Recently, there has been a lot of progress in showing fast convergence of LA even in cases where f is non-convex, notably \cite{VW19}, \cite{MoritaRisteski} in which the former paper focuses on functions f defined in \mathbb{R}^n and the latter paper focuses on functions with symmetries (like matrix completion type objectives) with manifold structure. Our work generalizes the results of \cite{VW19} where f is defined on a manifold M rather than \mathbb{R}^n . From technical point of view, we show that KL decreases in a geometric rate whenever the distribution e^{-f} satisfies a log-Sobolev inequality on M .

[Tensor Completion Made Practical](#)

- Allen Liu, Ankur Moitra
- abstract@[open-review](#): Tensor completion is a natural higher-order generalization of matrix completion where the goal is to recover a low-rank tensor from sparse observations of its entries. Existing algorithms are either heuristic without provable guarantees, based on solving large semidefinite programs which are impractical to run, or make strong assumptions such as requiring the factors to be nearly orthogonal. In this paper we introduce a new variant of alternating minimization, which in turn is inspired by understanding how the progress measures that guide convergence of alternating minimization in the matrix setting need to be adapted to the tensor setting. We show strong provable guarantees, including showing that our algorithm converges linearly to the true tensors even when the factors are highly correlated and can be implemented in nearly linear time. Moreover our algorithm is also highly practical and we show that we can complete third order tensors with a thousand dimensions from observing a tiny fraction of its entries. In contrast, and somewhat surprisingly, we show that the standard version of alternating minimization, without our new twist, can converge at a drastically slower rate in practice.

[Optimization and Generalization Analysis of Transduction through Gradient Boosting and Application to Multi-scale Graph Neural Networks](#)

- Kenta Oono, Taiji Suzuki
- abstract@[open-review](#): It is known that the current graph neural networks (GNNs) are difficult to make themselves deep due to the problem known as over-smoothing. Multi-scale GNNs are a promising approach for mitigating the over-smoothing problem. However, there is little explanation of why it works empirically from the viewpoint of learning theory. In this study, we derive the optimization and generalization guarantees of transductive learning algorithms that include multi-scale GNNs. Using the boosting theory, we prove the convergence of the training error under weak learning-type conditions. By combining it with generalization gap bounds in terms of transductive Rademacher complexity, we show that a test error bound of a specific type of multi-scale GNNs that decreases corresponding to the number of node aggregations under some conditions. Our results offer theoretical explanations for the effectiveness of the multi-scale structure against the over-smoothing problem. We apply boosting algorithms to the training of multi-scale GNNs for real-world node prediction tasks. We confirm that its performance is comparable to existing GNNs, and the practical behaviors are consistent with theoretical observations. Code is available at <https://github.com/delta2323/GB-GNN>.

[Content Provider Dynamics and Coordination in Recommendation Ecosystems](#)

- Omer Ben-Porat, Itay Rosenberg, Moshe Tennenholtz
- abstract@[open-review](#): Recommendation Systems like YouTube are vibrant ecosystems with two types of users: Content consumers (those who watch videos) and content providers (those who create videos). While the computational task of recommending relevant content is largely solved, designing a system that guarantees high social welfare for \textit{all} stakeholders is still in its infancy. In this work, we investigate the dynamics of content creation using a game-theoretic lens. Employing a stylized model that was recently suggested by other works, we show that the dynamics will always converge to a pure Nash Equilibrium (PNE), but the convergence rate can be exponential. We complement the analysis by proposing an efficient PNE computation algorithm via a combinatorial optimization problem that is of independent interest.

[Almost Surely Stable Deep Dynamics](#)

- Nathan Lawrence, Philip Loewen, Michael Forbes, Johan Backstrom, Bhushan Gopaluni
- abstract@[open-review](#): We introduce a method for learning provably stable deep neural network based dynamic models from observed data. Specifically, we consider discrete-time stochastic dynamic models, as they are of particular interest in practical applications such as estimation and control. However, these aspects exacerbate the challenge of guaranteeing stability. Our method works by embedding a Lyapunov neural network into the dynamic model, thereby inherently satisfying the stability criterion. To this end, we propose two approaches and apply them in both the deterministic and stochastic settings: one exploits convexity of the Lyapunov function, while the other enforces stability through an implicit output layer. We demonstrate the utility of each approach through numerical examples.

[Experimental design for MRI by greedy policy search](#)

- Tim Bakker, Herke van Hoof, Max Welling
- abstract@[open-review](#): In today's clinical practice, magnetic resonance imaging (MRI) is routinely accelerated through subsampling of the associated Fourier domain. Currently, the construction of these subsampling strategies - known as experimental design - relies primarily on heuristics. We propose to learn experimental design strategies for accelerated MRI with policy gradient methods. Unexpectedly, our experiments show that a simple greedy approximation of the objective leads to solutions nearly on-par with the more general non-greedy approach. We offer a partial explanation for this phenomenon rooted in greater variance in the non-greedy objective's gradient estimates, and experimentally verify that this variance hampers non-greedy models in adapting their policies to individual MR images. We empirically show that this adaptivity is key to improving subsampling designs.

[Expert-Supervised Reinforcement Learning for Offline Policy Learning and Evaluation](#)

- Aaron Sonabend, Junwei Lu, Leo Anthony Celi, Tianxi Cai, Peter Szolovits
- abstract@[open-review](#): Offline Reinforcement Learning (RL) is a promising approach for learning optimal policies in environments where direct exploration is expensive or unfeasible. However, the adoption of such policies in practice is often challenging, as they are hard to interpret within the application context, and lack measures of uncertainty for the learned policy value and its decisions. To overcome these issues, we propose an Expert-Supervised RL (ESRL) framework which uses uncertainty quantification for offline policy learning. In particular, we have three contributions: 1) the method can learn safe and optimal policies through hypothesis testing, 2) ESRL allows for different levels of risk averse implementations tailored to the application context, and finally, 3) we propose a way to interpret ESRL's policy at every state through posterior distributions, and use this framework to compute off-policy value function posteriors. We provide theoretical guarantees for our estimators and regret bounds consistent with Posterior Sampling for RL (PSRL). Sample efficiency of ESRL is independent of the chosen risk aversion threshold and quality of the behavior policy.

[ColdGANs: Taming Language GANs with Cautious Sampling Strategies](#)

- Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano
- abstract@[open-review](#): Training regimes based on Maximum Likelihood Estimation (MLE) suffer from known limitations, often leading to poorly generated text sequences that lack of coherence, factualness, and are prone to repetitions. At the root of these limitations is the mismatch between training and inference, i.e. the so-called exposure bias. Another problem lies in considering only the reference text as correct, while in practice several alternative

formulations could be as good. Generative Adversarial Networks (GANs) could mitigate those limitations. Nonetheless, the discrete nature of text has hindered their application to language generation: the approaches proposed so far, based on Reinforcement Learning, have been shown to under-perform MLE. In this context, the exploration is known to be critical, while surprisingly being under-studied. In this work, we show how the most popular sampling method results in unstable training for language GANs. We propose alternative exploration strategies that we named Cold-GANs. By forcing the sampling to be close to the distribution mode, the learning dynamic becomes smoother. We report experimental results obtained on three tasks: unconditional text generation, question generation, and abstractive summarization. For the first time, to the best of our knowledge, the proposed language GANs compare favorably to MLE, and obtain improvements over the state-of-the-art on the considered tasks.

[Hedging in games: Faster convergence of external and swap regrets](#)

- Xi Chen, Binghui Peng
- abstract@[open-review](#): We consider the setting where players run the Hedge algorithm or its optimistic variant \cite{syrgkanis2015fast} to play an n-action game repeatedly for T rounds. 1) For two-player games, we show that the regret of optimistic Hedge decays at $\tilde{O}(1/T^{5/6})$, improving the previous bound $O(1/T^{3/4})$ by \cite{syrgkanis2015fast}. 2) In contrast, we show that the convergence rate of vanilla Hedge is no better than $\tilde{O}(\sqrt{T})$, addressing an open question posted in \cite{syrgkanis2015fast}. For general m-player games, we show that the swap regret of each player decays at rate $\tilde{O}(m^{1/2}(n/T)^{3/4})$ when they combine optimistic Hedge with the classical external-to-internal reduction of Blum and Mansour \cite{blum2007external}. The algorithm can also be modified to achieve the same rate against itself and a rate of $\tilde{O}(\sqrt{n/T})$ against adversaries. Via standard connections, our upper bounds also imply faster convergence to coarse correlated equilibria in two-player games and to correlated equilibria in multiplayer games.

[The Origins and Prevalence of Texture Bias in Convolutional Neural Networks](#)

- Katherine Hermann, Ting Chen, Simon Kornblith
- abstract@[open-review](#): Recent work has indicated that, unlike humans, ImageNet-trained CNNs tend to classify images by texture rather than by shape. How pervasive is this bias, and where does it come from? We find that, when trained on datasets of images with conflicting shape and texture, CNNs learn to classify by shape at least as easily as by texture. What factors, then, produce the texture bias in CNNs trained on ImageNet? Different unsupervised training objectives and different architectures have small but significant and largely independent effects on the level of texture bias. However, all objectives and architectures still lead to models that make texture-based classification decisions a majority of the time, even if shape information is decodable from their hidden representations. The effect of data augmentation is much larger. By taking less aggressive random crops at training time and applying simple, naturalistic augmentation (color distortion, noise, and blur), we train models that classify ambiguous images by shape a majority of the time, and outperform baselines on out-of-distribution test sets. Our results indicate that apparent differences in the way humans and ImageNet-trained CNNs process images may arise not primarily from differences in their internal workings, but from differences in the data that they see.

[Time-Reversal Symmetric ODE Network](#)

- In Huh, Eunho Yang, Sung Ju Hwang, Jinwoo Shin
- abstract@[open-review](#): Time-reversal symmetry, which requires that the dynamics of a system should not change with the reversal of time axis, is a fundamental property that frequently holds in classical and quantum mechanics. In this paper, we propose a novel loss function that measures how well our ordinary differential equation (ODE) networks comply with this time-reversal symmetry; it is formally defined by the discrepancy in the time evolutions of ODE networks between forward and backward dynamics. Then, we design a new framework, which we name as Time-Reversal Symmetric ODE Networks (TRS-ODENs), that can learn the dynamics of physical systems more sample-efficiently by learning with the proposed loss function. We evaluate TRS-ODENs on several classical dynamics, and find they can learn the desired time evolution from observed noisy and complex trajectories. We also show that, even for systems that do not possess the full time-reversal symmetry, TRS-ODENs can achieve better predictive performances over baselines.

[Provable Overlapping Community Detection in Weighted Graphs](#)

- Jimit Majmudar, Stephen Vavasis
- abstract@[open-review](#): Community detection is a widely-studied unsupervised learning problem in which the task is to group similar entities together based on observed pairwise entity interactions. This problem has applications in diverse domains such as social network analysis and computational biology. There is a significant amount of literature studying this problem under the assumption that the communities do not overlap. When the communities are allowed to overlap, often a \textit{pure nodes} assumption is made, i.e. each community has a node that belongs exclusively to that community. This assumption, however, may not always be satisfied in practice. In this paper, we provide a provable method to detect overlapping communities in weighted graphs without explicitly making the pure nodes assumption. Moreover, contrary to most existing algorithms, our approach is based on convex optimization, for which many useful theoretical properties are already known. We demonstrate the success of our algorithm on artificial and real-world datasets.

[Fast Unbalanced Optimal Transport on a Tree](#)

- Ryoma Sato, Makoto Yamada, Hisashi Kashima
- abstract@[open-review](#): This study examines the time complexities of the unbalanced optimal transport problems from an algorithmic perspective for the first time. We reveal which problems in unbalanced optimal transport can/cannot be solved efficiently. Specifically, we prove that the Kantorovich Rubinstein distance and optimal partial transport in Euclidean metric cannot be computed in strongly subquadratic time under the strong exponential time hypothesis. Then, we propose an algorithm that solves a more general unbalanced optimal transport problem exactly in quasi-linear time on a tree metric. The proposed algorithm processes a tree with one million nodes in less than one second. Our analysis forms a foundation for the theoretical study of unbalanced optimal transport algorithms and opens the door to the applications of unbalanced optimal transport to million-scale datasets.

[Acceleration with a Ball Optimization Oracle](#)

- Yair Carmon, Arun Jambulapati, Qijia Jiang, Yujia Jin, Yin Tat Lee, Aaron Sidford, Kevin Tian
- abstract@[open-review](#): Consider an oracle which takes a point x and returns the minimizer of a convex function f in an L_2 ball of radius r around x . It is straightforward to show that roughly $r^{-1}\log(1/\epsilon)$ calls to the oracle suffice to find an ϵ -approximate minimizer of f in an L_2 unit ball. Perhaps surprisingly, this is not optimal: we design an accelerated algorithm which attains an ϵ -approximate minimizer with roughly $r^{-2/3}\log(1/\epsilon)$ oracle queries, and give a matching lower bound. Further, we implement ball optimization oracles for functions with a locally stable Hessian using a variant of Newton's method and, in certain cases, stochastic first-order methods. The resulting algorithms apply to a number of problems of practical and theoretical import, improving upon previous results for logistic and linfinity regression and achieving guarantees comparable to the state-of-the-art for L_p regression.

[Avoiding Side Effects By Considering Future Tasks](#)

- Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, Shane Legg

- abstract@[open-review](#): Designing reward functions is difficult: the designer has to specify what to do (what it means to complete the task) as well as what not to do (side effects that should be avoided while completing the task). To alleviate the burden on the reward designer, we propose an algorithm to automatically generate an auxiliary reward function that penalizes side effects. This auxiliary objective rewards the ability to complete possible future tasks, which decreases if the agent causes side effects during the current task. The future task reward can also give the agent an incentive to interfere with events in the environment that make future tasks less achievable, such as irreversible actions by other agents. To avoid this interference incentive, we introduce a baseline policy that represents a default course of action (such as doing nothing), and use it to filter out future tasks that are not achievable by default. We formally define interference incentives and show that the future task approach with a baseline policy avoids these incentives in the deterministic case. Using gridworld environments that test for side effects and interference, we show that our method avoids interference and is more effective for avoiding side effects than the common approach of penalizing irreversible actions.

[Handling Missing Data with Graph Representation Learning](#)

- Jiaxuan You, Xiaobai Ma, Yi Ding, Mykel J. Kochenderfer, Jure Leskovec
- abstract@[open-review](#): Machine learning with missing data has been approached in many different ways, including feature imputation where missing feature values are estimated based on observed values and label prediction where downstream labels are learned directly from incomplete data. However, existing imputation models tend to have strong prior assumptions and cannot learn from downstream tasks, while models targeting label predictions often involve heuristics and can encounter scalability issues. Here we propose GRAPE, a framework for feature imputation as well as label prediction. GRAPE tackles the missing data problem using graph representation, where the observations and features are viewed as two types of nodes in a bipartite graph, and the observed feature values as edges. Under the GRAPE framework, the feature imputation is formulated as an edge-level prediction task and the label prediction as a node-level prediction task. These tasks are then solved with Graph Neural Networks. Experimental results on nine benchmark datasets show that GRAPE yields 20% lower mean absolute error for imputation tasks and 10% lower for label prediction tasks, compared with existing state-of-the-art methods.

[Improving Auto-Augment via Augmentation-Wise Weight Sharing](#)

- Keyu Tian, Chen Lin, Ming Sun, Luping Zhou, Junjie Yan, Wanli Ouyang
- abstract@[open-review](#): The recent progress on automatically searching augmentation policies has boosted the performance substantially for various tasks. A key component of automatic augmentation search is the evaluation process for a particular augmentation policy, which is utilized to return reward and usually runs thousands of times. A plain evaluation process, which includes full model training and validation, would be time-consuming. To achieve efficiency, many choose to sacrifice evaluation reliability for speed. In this paper, we dive into the dynamics of augmented training of the model. This inspires us to design a powerful and efficient proxy task based on the Augmentation-Wise Weight Sharing (AWS) to form a fast yet accurate evaluation process in an elegant way. Comprehensive analysis verifies the superiority of this approach in terms of effectiveness and efficiency. The augmentation policies found by our method achieve superior accuracies compared with existing auto-augmentation search methods. On CIFAR-10, we achieve a top-1 error rate of 1.24%, which is currently the best performing single model without extra training data. On ImageNet, we get a top-1 error rate of 20.36% for ResNet-50, which leads to 3.34% absolute error rate reduction over the baseline augmentation.

[MMA Regularization: Decorrelating Weights of Neural Networks by Maximizing the Minimal Angles](#)

- Zhennan Wang, Canqun Xiang, Wenbin Zou, Chen Xu
- abstract@[open-review](#): The strong correlation between neurons or filters can significantly weaken the generalization ability of neural networks. Inspired by the well-known Tammes problem, we propose a novel diversity regularization method to address this issue, which makes the normalized weight vectors of neurons or filters distributed on a hypersphere as uniformly as possible, through maximizing the minimal pairwise angles (MMA). This method can easily exert its effect by plugging the MMA regularization term into the loss function with negligible computational overhead. The MMA regularization is simple, efficient, and effective. Therefore, it can be used as a basic regularization method in neural network training. Extensive experiments demonstrate that MMA regularization is able to enhance the generalization ability of various modern models and achieves considerable performance improvements on CIFAR100 and TinyImageNet datasets. In addition, experiments on face verification show that MMA regularization is also effective for feature learning. Code is available at: <https://github.com/wznpub/MMA-Regularization>.

[HRN: A Holistic Approach to One Class Learning](#)

- Wenpeng Hu, Mengyu Wang, Qi Qin, Jinwen Ma, Bing Liu
- abstract@[open-review](#): Existing neural network based one-class learning methods mainly use various forms of auto-encoders or GAN style adversarial training to learn a latent representation of the given one class of data. This paper proposes an entirely different approach based on a novel regularization, called holistic regularization (or H-regularization), which enables the system to consider the data holistically, not to produce a model that biases towards some features. Combined with a proposed 2-norm instance-level data normalization, we obtain an effective one-class learning method, called HRN. To our knowledge, the proposed regularization and the normalization method have not been reported before. Experimental evaluation using both benchmark image classification and traditional anomaly detection datasets show that HRN markedly outperforms the state-of-the-art existing deep/non-deep learning models.

[The Generalized Lasso with Nonlinear Observations and Generative Priors](#)

- Zhaoqiang Liu, Jonathan Scarlett
- abstract@[open-review](#): In this paper, we study the problem of signal estimation from noisy non-linear measurements when the unknown $\$n\$$ -dimensional signal is in the range of an $\$L\$$ -Lipschitz continuous generative model with bounded $\$k\$$ -dimensional inputs. We make the assumption of sub-Gaussian measurements, which is satisfied by a wide range of measurement models, such as linear, logistic, 1-bit, and other quantized models. In addition, we consider the impact of adversarial corruptions on these measurements. Our analysis is based on a generalized Lasso approach (Plan and Vershynin, 2016). We first provide a non-uniform recovery guarantee, which states that under i.i.d. \sim Gaussian measurements, roughly $\$O(\frac{1}{\epsilon^2} \log L)^{\frac{1}{2}}$ samples suffice for recovery with an $\|\cdot\|_2$ -error of ϵ , and that this scheme is robust to adversarial noise. Then, we apply this result to neural network generative models, and discuss various extensions to other models and non-i.i.d. \sim measurements. Moreover, we show that our result can be extended to the uniform recovery guarantee under the assumption of a so-called local embedding property, which is satisfied by the 1-bit and censored Tobit models.

[Fair regression via plug-in estimator and recalibration with statistical guarantees](#)

- Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, Massimiliano Pontil
- abstract@[open-review](#): We study the problem of learning an optimal regression function subject to a fairness constraint. It requires that, conditionally on the sensitive feature, the distribution of the function output remains the same. This constraint naturally extends the notion of demographic parity, often used in classification, to the regression setting. We tackle this problem by leveraging on a proxy-discretized version, for which we derive an explicit expression of the optimal fair predictor. This result naturally suggests a two stage approach, in which we first estimate the (unconstrained) regression function from a set of labeled data and then we recalibrate it with another set of unlabeled data. The recalibration step can be efficiently performed via a smooth optimization. We derive rates of convergence of the proposed estimator to the optimal fair predictor both in terms of the risk and fairness constraint. Finally, we present numerical experiments illustrating that the proposed method is often superior or competitive with state-of-the-art methods.

[Modeling Shared responses in Neuroimaging Studies through MultiView ICA](#)

- Hugo Richard, Luigi Gresele, Aapo Hyvärinen, Bertrand Thirion, Alexandre Gramfort, Pierre Ablin
- abstract@[open-review](#): Group studies involving large cohorts of subjects are important to draw general conclusions about brain functional organization. However, the aggregation of data coming from multiple subjects is challenging, since it requires accounting for large variability in anatomy, functional topography and stimulus response across individuals. Data modeling is especially hard for ecologically relevant conditions such as movie watching, where the experimental setup does not imply well-defined cognitive operations. We propose a novel MultiView Independent Component Analysis (ICA) model for group studies, where data from each subject are modeled as a linear combination of shared independent sources plus noise. Contrary to most group-ICA procedures, the likelihood of the model is available in closed form. We develop an alternate quasi-Newton method for maximizing the likelihood, which is robust and converges quickly. We demonstrate the usefulness of our approach first on fMRI data, where our model demonstrates improved sensitivity in identifying common sources among subjects. Moreover, the sources recovered by our model exhibit lower between-sessions variability than other methods. On magnetoencephalography (MEG) data, our method yields more accurate source localization on phantom data. Applied on 200 subjects from the Cam-CAN dataset, it reveals a clear sequence of evoked activity in sensor and source space.

[Efficient Planning in Large MDPs with Weak Linear Function Approximation](#)

- Roshan Shariff, Csaba Szepesvari
- abstract@[open-review](#): Large-scale Markov decision processes (MDPs) require planning algorithms with runtime independent of the number of states of the MDP. We consider the planning problem in MDPs using linear value function approximation with only weak requirements: low approximation error for the optimal value function, and a small set of “core” states whose features span those of other states. In particular, we make no assumptions about the representability of policies or value functions of non-optimal policies. Our algorithm produces almost-optimal actions for any state using a generative oracle (simulator) for the MDP, while its computation time scales polynomially with the number of features, core states, and actions and the effective horizon.

[Efficient Learning of Generative Models via Finite-Difference Score Matching](#)

- Tianyu Pang, Kun Xu, Chongxuan LI, Yang Song, Stefano Ermon, Jun Zhu
- abstract@[open-review](#): Several machine learning applications involve the optimization of higher-order derivatives (e.g., gradients of gradients) during training, which can be expensive with respect to memory and computation even with automatic differentiation. As a typical example in generative modeling, score matching (SM) involves the optimization of the trace of a Hessian. To improve computing efficiency, we rewrite the SM objective and its variants in terms of directional derivatives, and present a generic strategy to efficiently approximate any-order directional derivative with finite difference (FD). Our approximation only involves function evaluations, which can be executed in parallel, and no gradient computations. Thus, it reduces the total computational cost while also improving numerical stability. We provide two instantiations by reformulating variants of SM objectives into the FD forms. Empirically, we demonstrate that our methods produce results comparable to the gradient-based counterparts while being much more computationally efficient.

[Semialgebraic Optimization for Lipschitz Constants of ReLU Networks](#)

- Tong Chen, Jean B. Lasserre, Victor Magron, Edouard Pauwels
- abstract@[open-review](#): The Lipschitz constant of a network plays an important role in many applications of deep learning, such as robustness certification and Wasserstein Generative Adversarial Network. We introduce a semidefinite programming hierarchy to estimate the global and local Lipschitz constant of a multiple layer deep neural network. The novelty is to combine a polynomial lifting for ReLU functions derivatives with a weak generalization of Putinar's positivity certificate. This idea could also apply to other, nearly sparse, polynomial optimization problems in machine learning. We empirically demonstrate that our method provides a trade-off with respect to state of the art linear programming approach, and in some cases we obtain better bounds in less time.

[Linear-Sample Learning of Low-Rank Distributions](#)

- Ayush Jain, Alon Orlitsky
- abstract@[open-review](#): Many latent-variable applications, including community detection, collaborative filtering, genomic analysis, and NLP, model data as generated by low-rank matrices. Yet despite considerable research, except for very special cases, the number of samples required to efficiently recover the underlying matrices has not been known. We determine the onset of learning in several common latent-variable settings. For all of them, we show that learning $k \times k$, rank- r , matrices to normalized L_1 distance ϵ requires $\Omega(\frac{kr}{\epsilon^2})$ samples, and propose an algorithm that uses $O(\frac{kr}{\epsilon^2} \log^2 r/\epsilon)$ samples, a number linear in the high dimension, and nearly linear in the, typically low, rank. The algorithm improves on existing spectral techniques and runs in polynomial time. The proofs establish new results on the rapid convergence of the spectral distance between the model and observation matrices, and may be of independent interest.

[Transferable Calibration with Lower Bias and Variance in Domain Adaptation](#)

- Ximei Wang, Mingsheng Long, Jianmin Wang, Michael Jordan
- abstract@[open-review](#): Domain Adaptation (DA) enables transferring a learning machine from a labeled source domain to an unlabeled target one. While remarkable advances have been made, most of the existing DA methods focus on improving the target accuracy at inference. How to estimate the predictive uncertainty of DA models is vital for decision-making in safety-critical scenarios but remains the boundary to explore. In this paper, we delve into the open problem of Calibration in DA, which is extremely challenging due to the coexistence of domain shift and the lack of target labels. We first reveal the dilemma that DA models learn higher accuracy at the expense of well-calibrated probabilities. Driven by this finding, we propose Transferable Calibration (TransCal) to achieve more accurate calibration with lower bias and variance in a unified hyperparameter-free optimization framework. As a general post-hoc calibration method, TransCal can be easily applied to recalibrate existing DA methods. Its efficacy has been justified both theoretically and empirically.

[Generalization bound of globally optimal non-convex neural network training: Transportation map estimation by infinite dimensional Langevin dynamics](#)

- Taiji Suzuki
- abstract@[open-review](#): We introduce a new theoretical framework to analyze deep learning optimization with connection to its generalization error. Existing frameworks such as mean field theory and neural tangent kernel theory for neural network optimization analysis typically require taking limit of infinite width of the network to show its global convergence. This potentially makes it difficult to directly deal with finite width network; especially in the neural tangent kernel regime, we cannot reveal favorable properties of neural networks {it beyond kernel methods}. To realize more natural analysis, we consider a completely different approach in which we formulate the parameter training as a transportation map estimation and show its global convergence via the theory of the {it infinite dimensional Langevin dynamics}. This enables us to analyze narrow and wide networks in a unifying manner. Moreover, we give generalization gap and excess risk bounds for the solution obtained by the dynamics. The excess risk bound achieves the so-called fast learning rate. In particular, we show an exponential convergence for a classification problem and a minimax optimal rate for a regression problem.

[Online Bayesian Goal Inference for Boundedly Rational Planning Agents](#)

- Tan Zhi-Xuan, Jordyn Mann, Tom Silver, Josh Tenenbaum, Vikash Mansinghka
- abstract@[open-review](#): People routinely infer the goals of others by observing their actions over time. Remarkably, we can do so even when those actions lead to failure, enabling us to assist others when we detect that they might not achieve their goals. How might we endow machines with similar capabilities? Here we present an architecture capable of inferring an agent's goals online from both optimal and non-optimal sequences of actions. Our architecture models agents as boundedly-rational planners that interleave search with execution by replanning, thereby accounting for sub-optimal behavior. These models are specified as probabilistic programs, allowing us to represent and perform efficient Bayesian inference over an agent's goals and internal planning processes. To perform such inference, we develop Sequential Inverse Plan Search (SIPS), a sequential Monte Carlo algorithm that exploits the online replanning assumption of these models, limiting computation by incrementally extending inferred plans as new actions are observed. We present experiments showing that this modeling and inference architecture outperforms Bayesian inverse reinforcement learning baselines, accurately inferring goals from both optimal and non-optimal trajectories involving failure and back-tracking, while generalizing across domains with compositional structure and sparse rewards.

[BayReL: Bayesian Relational Learning for Multi-omics Data Integration](#)

- Ehsan Hajiramezanali, Arman Hasanzadeh, Nick Duffield, Krishna Narayanan, Xiaoning Qian
- abstract@[open-review](#): High-throughput molecular profiling technologies have produced high-dimensional multi-omics data, enabling systematic understanding of living systems at the genome scale. Studying molecular interactions across different data types helps reveal signal transduction mechanisms across different classes of molecules. In this paper, we develop a novel Bayesian representation learning method that infers the relational interactions across multi-omics data types. Our method, Bayesian Relational Learning (BayReL) for multi-omics data integration, takes advantage of a priori known relationships among the same class of molecules, modeled as a graph at each corresponding view, to learn view-specific latent variables as well as a multi-partite graph that encodes the interactions across views. Our experiments on several real-world datasets demonstrate enhanced performance of BayReL in inferring meaningful interactions compared to existing baselines.

[Weakly Supervised Deep Functional Maps for Shape Matching](#)

- Abhishek Sharma, Maks Ovsjanikov
- abstract@[open-review](#): A variety of deep functional maps have been proposed recently, from fully supervised to totally unsupervised, with a range of loss functions as well as different regularization terms. However, it is still not clear what are minimum ingredients of a deep functional map pipeline and whether such ingredients unify or generalize all recent work on deep functional maps. We show empirically the minimum components for obtaining state-of-the-art results with different loss functions, supervised as well as unsupervised. Furthermore, we propose a novel framework designed for both full-to-full as well as partial to full shape matching that achieves state of the art results on several benchmark datasets outperforming, even the fully supervised methods. Our code is publicly available at \url{https://github.com/Not-IITian/Weakly-supervised-Functional-map}

[Domain Adaptation with Conditional Distribution Matching and Generalized Label Shift](#)

- Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, Geoffrey J. Gordon
- abstract@[open-review](#): Adversarial learning has demonstrated good performance in the unsupervised domain adaptation setting, by learning domain-invariant representations. However, recent work has shown limitations of this approach when label distributions differ between the source and target domains. In this paper, we propose a new assumption, \textit{generalized label shift} (\textit{glsa}), to improve robustness against mismatched label distributions. \textit{glsa} states that, conditioned on the label, there exists a representation of the input that is invariant between the source and target domains. Under \textit{glsa}, we provide theoretical guarantees on the transfer performance of any classifier. We also devise necessary and sufficient conditions for \textit{glsa} to hold, by using an estimation of the relative class weights between domains and an appropriate reweighting of samples. Our weight estimation method could be straightforwardly and generically applied in existing domain adaptation (DA) algorithms that learn domain-invariant representations, with small computational overhead. In particular, we modify three DA algorithms, JAN, DANN and CDAN, and evaluate their performance on standard and artificial DA tasks. Our algorithms outperform the base versions, with vast improvements for large label distribution mismatches. Our code is available at \url{https://tinyurl.com/y585xt6j}.

[Rethinking the Value of Labels for Improving Class-Imbalanced Learning](#)

- Yuzhe Yang, Zhi Xu
- abstract@[open-review](#): Real-world data often exhibits long-tailed distributions with heavy class imbalance, posing great challenges for deep recognition models. We identify a persisting dilemma on the value of labels in the context of imbalanced learning: on the one hand, supervision from labels typically leads to better results than its unsupervised counterparts; on the other hand, heavily imbalanced data naturally incurs "label bias" in the classifier, where the decision boundary can be drastically altered by the majority classes. In this work, we systematically investigate these two facets of labels. We demonstrate, theoretically and empirically, that class-imbalanced learning can significantly benefit in both semi-supervised and self-supervised manners. Specifically, we confirm that (1) positively, imbalanced labels are valuable: given more unlabeled data, the original labels can be leveraged with the extra data to reduce label bias in a semi-supervised manner, which greatly improves the final classifier; (2) negatively however, we argue that imbalanced labels are not useful always: classifiers that are first pre-trained in a self-supervised manner consistently outperform their corresponding baselines. Extensive experiments on large-scale imbalanced datasets verify our theoretically grounded strategies, showing superior performance over previous state-of-the-arts. Our intriguing findings highlight the need to rethink the usage of imbalanced labels in realistic long-tailed tasks. Code is available at <https://github.com/YyzHarry/imbalanced-semi-self>.

[Provably Robust Metric Learning](#)

- Lu Wang, Xuanqing Liu, Jinfeng Yi, Yuan Jiang, Cho-Jui Hsieh
- abstract@[open-review](#): Metric learning is an important family of algorithms for classification and similarity search, but the robustness of learned metrics against small adversarial perturbations is less studied. In this paper, we show that existing metric learning algorithms, which focus on boosting the clean accuracy, can result in metrics that are less robust than the Euclidean distance. To overcome this problem, we propose a novel metric learning algorithm to find a Mahalanobis distance that is robust against adversarial perturbations, and the robustness of the resulting model is certifiable. Experimental results show that the proposed metric learning algorithm improves both certified robust errors and empirical robust errors (errors under adversarial attacks). Furthermore, unlike neural network defenses which usually encounter a trade-off between clean and robust errors, our method does not sacrifice clean errors compared with previous metric learning methods.

[Iterative Deep Graph Learning for Graph Neural Networks: Better and Robust Node Embeddings](#)

- Yu Chen, Lingfei Wu, Mohammed Zaki
- abstract@[open-review](#): In this paper, we propose an end-to-end graph learning framework, namely \textbf{Iterative Deep Graph Learning} (\textbf{alg}), for jointly and iteratively learning graph structure and graph embedding. The key rationale of \textbf{alg} is to learn a better graph structure based on better node embeddings, and vice versa (i.e., better node embeddings based on a better graph structure). Our iterative method

dynamically stops when the learned graph structure approaches close enough to the graph optimized for the downstream prediction task. In addition, we cast the graph learning problem as a similarity metric learning problem and leverage adaptive graph regularization for controlling the quality of the learned graph. Finally, combining the anchor-based approximation technique, we further propose a scalable version of \alg, namely \salg, which significantly reduces the time and space complexity of \alg without compromising the performance. Our extensive experiments on nine benchmarks show that our proposed \alg models can consistently outperform or match the state-of-the-art baselines. Furthermore, \alg can be more robust to adversarial graphs and cope with both transductive and inductive learning.

[COPT: Coordinated Optimal Transport on Graphs](#)

- Yihe Dong, Will Sawin
- abstract@[open-review](#): We introduce COPT, a novel distance metric between graphs defined via an optimization routine, computing a coordinated pair of optimal transport maps simultaneously. This gives an unsupervised way to learn general-purpose graph representation, applicable to both graph sketching and graph comparison. COPT involves simultaneously optimizing dual transport plans, one between the vertices of two graphs, and another between graph signal probability distributions. We show theoretically that our method preserves important global structural information on graphs, in particular spectral information, and analyze connections to existing studies. Empirically, COPT outperforms state of the art methods in graph classification on both synthetic and real datasets.

[No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems](#)

- Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, Christopher RÃ©
- abstract@[open-review](#): In real-world classification tasks, each class often comprises multiple finer-grained "subclasses." As the subclass labels are frequently unavailable, models trained using only the coarser-grained class labels often exhibit highly variable performance across different subclasses. This phenomenon, known as hidden stratification, has important consequences for models deployed in safety-critical applications such as medicine. We propose GEORGE, a method to both measure and mitigate hidden stratification even when subclass labels are unknown. We first observe that unlabeled subclasses are often separable in the feature space of deep models, and exploit this fact to estimate subclass labels for the training data via clustering techniques. We then use these approximate subclass labels as a form of noisy supervision in a distributionally robust optimization objective. We theoretically characterize the performance of GEORGE in terms of the worst-case generalization error across any subclass. We empirically validate GEORGE on a mix of real-world and benchmark image classification datasets, and show that our approach boosts worst-case subclass accuracy by up to 15 percentage points compared to standard training techniques, without requiring any information about the subclasses.

[Model Rubikâ€™s Cube: Twisting Resolution, Depth and Width for TinyNets](#)

- Kai Han, Yunhe Wang, Qiulin Zhang, Wei Zhang, Chunjing XU, Tong Zhang
- abstract@[open-review](#): To obtain excellent deep neural architectures, a series of techniques are carefully designed in EfficientNets. The giant formula for simultaneously enlarging the resolution, depth and width provides us a Rubikâ€™s cube for neural networks. So that we can find networks with high efficiency and excellent performance by twisting the three dimensions. This paper aims to explore the twisting rules for obtaining deep neural networks with minimum model sizes and computational costs. Different from the network enlarging, we observe that resolution and depth are more important than width for tiny networks. Therefore, the original method, i.e the compound scaling in EfficientNet is no longer suitable. To this end, we summarize a tiny formula for downsizing neural architectures through a series of smaller models derived from the EfficientNet-B0 with the FLOPs constraint. Experimental results on the ImageNet benchmark illustrate that our TinyNet performs much better than the smaller version of EfficientNets using the inverted giant formula. For instance, our TinyNet-E achieves a 59.9% Top-1 accuracy with only 24M FLOPs, which is about 1.9% higher than that of the previous best MobileNetV3 with similar computational cost. Code will be available at \url{<https://github.com/huawei-noah/CV-Backbones/tree/master/tinynet>}, and \url{https://gitee.com/mindspore/mindspore/tree/master/model_zoo/research/cv/tinynet}.

[Self-Adaptive Training: beyond Empirical Risk Minimization](#)

- Lang Huang, Chao Zhang, Hongyang Zhang
- abstract@[open-review](#): We propose self-adaptive training---a new training algorithm that dynamically calibrates training process by model predictions without incurring extra computational cost---to improve generalization of deep learning for potentially corrupted training data. This problem is important to robustly learning from data that are corrupted by, e.g., random noises and adversarial examples. The standard empirical risk minimization (ERM) for such data, however, may easily overfit noises and thus suffers from sub-optimal performance. In this paper, we observe that model predictions can substantially benefit the training process: self-adaptive training significantly mitigates the overfitting issue and improves generalization over ERM under both random and adversarial noises. Besides, in sharp contrast to the recently-discovered double-descent phenomenon in ERM, self-adaptive training exhibits a single-descent error-capacity curve, indicating that such a phenomenon might be a result of overfitting of noises. Experiments on the CIFAR and ImageNet datasets verify the effectiveness of our approach in two applications: classification with label noise and selective classification.

[Effective Dimension Adaptive Sketching Methods for Faster Regularized Least-Squares Optimization](#)

- Jonathan Lacotte, Mert Pilanci
- abstract@[open-review](#): We propose a new randomized algorithm for solving L2-regularized least-squares problems based on sketching. We consider two of the most popular random embeddings, namely, Gaussian embeddings and the Subsampled Randomized Hadamard Transform (SRHT). While current randomized solvers for least-squares optimization prescribe an embedding dimension at least greater than the data dimension, we show that the embedding dimension can be reduced to the effective dimension of the optimization problem, and still preserve high-probability convergence guarantees. In this regard, we derive sharp matrix deviation inequalities over ellipsoids for both Gaussian and SRHT embeddings. Specifically, we improve on the constant of a classical Gaussian concentration bound whereas, for SRHT embeddings, our deviation inequality involves a novel technical approach. Leveraging these bounds, we are able to design a practical and adaptive algorithm which does not require to know the effective dimension beforehand. Our method starts with an initial embedding dimension equal to 1 and, over iterations, increases the embedding dimension up to the effective one at most. Hence, our algorithm improves the state-of-the-art computational complexity for solving regularized least-squares problems. Further, we show numerically that it outperforms standard iterative solvers such as the conjugate gradient method and its pre-conditioned version on several standard machine learning datasets.

[Near-Optimal Comparison Based Clustering](#)

- MichaÃ«l Perrot, Pascal Esser, Debarghya Ghoshdastidar
- abstract@[open-review](#): The goal of clustering is to group similar objects into meaningful partitions. This process is well understood when an explicit similarity measure between the objects is given. However, far less is known when this information is not readily available and, instead, one only observes ordinal comparisons such as ``object i is more similar to j than to k.'' In this paper, we tackle this problem using a two-step procedure: we estimate a pairwise similarity matrix from the comparisons before using a clustering method based on semi-definite programming (SDP). We theoretically show that our approach can exactly recover a planted clustering using a near-optimal number of passive comparisons. We empirically validate our theoretical findings and demonstrate the good behaviour of our method on real data.

[Multi-Task Temporal Shift Attention Networks for On-Device Contactless Vitals Measurement](#)

- Xin Liu, Josh Fromm, Shwetak Patel, Daniel McDuff
- abstract@[open-review](#): Telehealth and remote health monitoring have become increasingly important during the SARS-CoV-2 pandemic and it is widely expected that this will have a lasting impact on healthcare practices. These tools can help reduce the risk of exposing patients and medical staff to infection, make healthcare services more accessible, and allow providers to see more patients. However, objective measurement of vital signs is challenging without direct contact with a patient. We present a video-based and on-device optical cardiopulmonary vital sign measurement approach. It leverages a novel multi-task temporal shift convolutional attention network (MTTS-CAN) and enables real-time cardiovascular and respiratory measurements on mobile platforms. We evaluate our system on an Advanced RISC Machine (ARM) CPU and achieve state-of-the-art accuracy while running at over 150 frames per second which enables real-time applications. Systematic experimentation on large benchmark datasets reveals that our approach leads to substantial (20%-50%) reductions in error and generalizes well across datasets.

[A new convergent variant of Q-learning with linear function approximation](#)

- Diogo Carvalho, Francisco S. Melo, Pedro Santos
- abstract@[open-review](#): In this work, we identify a novel set of conditions that ensure convergence with probability 1 of Q-learning with linear function approximation, by proposing a two time-scale variation thereof. In the faster time scale, the algorithm features an update similar to that of DQN, where the impact of bootstrapping is attenuated by using a Q-value estimate akin to that of the target network in DQN. The slower time-scale, in turn, can be seen as a modified target network update. We establish the convergence of our algorithm, provide an error bound and discuss our results in light of existing convergence results on reinforcement learning with function approximation. Finally, we illustrate the convergent behavior of our method in domains where standard Q-learning has previously been shown to diverge.

[TaylorGAN: Neighbor-Augmented Policy Update Towards Sample-Efficient Natural Language Generation](#)

- Chun-Hsing Lin, Siang-Ruei Wu, Hung-yi Lee, Yun-Nung Chen
- abstract@[open-review](#): Score function-based natural language generation (NLG) approaches such as REINFORCE, in general, suffer from low sample efficiency and training instability problems. This is mainly due to the non-differentiable nature of the discrete space sampling and thus these methods have to treat the discriminator as a black box and ignore the gradient information. To improve the sample efficiency and reduce the variance of REINFORCE, we propose a novel approach, TaylorGAN, which augments the gradient estimation by off-policy update and the first-order Taylor expansion. This approach enables us to train NLG models from scratch with smaller batch size --- without maximum likelihood pre-training, and outperforms existing GAN-based methods on multiple metrics of quality and diversity.

[Neural Networks with Small Weights and Depth-Separation Barriers](#)

- Gal Vardi, Ohad Shamir
- abstract@[open-review](#): In studying the expressiveness of neural networks, an important question is whether there are functions which can only be approximated by sufficiently deep networks, assuming their size is bounded. However, for constant depths, existing results are limited to depths \$2\$ and \$3\$, and achieving results for higher depths has been an important open question. In this paper, we focus on feedforward ReLU networks, and prove fundamental barriers to proving such results beyond depth \$4\$, by reduction to open problems and natural-proof barriers in circuit complexity. To show this, we study a seemingly unrelated problem of independent interest: Namely, whether there are polynomially-bounded functions which require super-polynomial weights in order to approximate with constant-depth neural networks. We provide a negative and constructive answer to that question, by showing that if a function can be approximated by a polynomially-sized, constant depth \$k\$ network with arbitrarily large weights, it can also be approximated by a polynomially-sized, depth \$3k+3\$ network, whose weights are polynomially bounded.

[Untangling tradeoffs between recurrence and self-attention in artificial neural networks](#)

- Giancarlo Kerg, Bhargav Kanuparthi, Anirudh Goyal ALIAS PARTH GOYAL, Kyle Goyette, Yoshua Bengio, Guillaume Lajoie
- abstract@[open-review](#): Attention and self-attention mechanisms, are now central to state-of-the-art deep learning on sequential tasks. However, most recent progress hinges on heuristic approaches with limited understanding of attention's role in model optimization and computation, and rely on considerable memory and computational resources that scale poorly. In this work, we present a formal analysis of how self-attention affects gradient propagation in recurrent networks, and prove that it mitigates the problem of vanishing gradients when trying to capture long-term dependencies by establishing concrete bounds for gradient norms. Building on these results, we propose a relevancy screening mechanism, inspired by the cognitive process of memory consolidation, that allows for a scalable use of sparse self-attention with recurrence. While providing guarantees to avoid vanishing gradients, we use simple numerical experiments to demonstrate the tradeoffs in performance and computational resources by efficiently balancing attention and recurrence. Based on our results, we propose a concrete direction of research to improve scalability of attentive networks.

[Dual-Free Stochastic Decentralized Optimization with Variance Reduction](#)

- Hadrien Hendrikx, Francis Bach, Laurent MassouliÃ©
- abstract@[open-review](#): We consider the problem of training machine learning models on distributed data in a decentralized way. For finite-sum problems, fast single-machine algorithms for large datasets rely on stochastic updates combined with variance reduction. Yet, existing decentralized stochastic algorithms either do not obtain the full speedup allowed by stochastic updates, or require oracles that are more expensive than regular gradients. In this work, we introduce a Decentralized stochastic algorithm with Variance Reduction called DVR. DVR only requires computing stochastic gradients of the local functions, and is computationally as fast as a standard stochastic variance-reduced algorithms run on a \$1/n\$ fraction of the dataset, where \$n\$ is the number of nodes. To derive DVR, we use Bregman coordinate descent on a well-chosen dual problem, and obtain a dual-free algorithm using a specific Bregman divergence. We give an accelerated version of DVR based on the Catalyst framework, and illustrate its effectiveness with simulations on real data.

[Online Learning in Contextual Bandits using Gated Linear Networks](#)

- Eren Sezener, Marcus Hutter, David Budden, Jianan Wang, Joel Veness
- abstract@[open-review](#): We introduce a new and completely online contextual bandit algorithm called Gated Linear Contextual Bandits (GLCB). This algorithm is based on Gated Linear Networks (GLNs), a recently introduced deep learning architecture with properties well-suited to the online setting. Leveraging data-dependent gating properties of the GLN we are able to estimate prediction uncertainty with effectively zero algorithmic overhead. We empirically evaluate GLCB compared to 9 state-of-the-art algorithms that leverage deep neural networks, on a standard benchmark suite of discrete and continuous contextual bandit problems. GLCB obtains mean first-place despite being the only online method, and we further support these results with a theoretical study of its convergence properties.

[Throughput-Optimal Topology Design for Cross-Silo Federated Learning](#)

- Othmane MARFOQ, CHUAN XU, Giovanni Neglia, Richard Vidal

- abstract@[open-review](#): Federated learning usually employs a client-server architecture where an orchestrator iteratively aggregates model updates from remote clients and pushes them back a refined model. This approach may be inefficient in cross-silo settings, as close-by data silos with high-speed access links may exchange information faster than with the orchestrator, and the orchestrator may become a communication bottleneck. In this paper we define the problem of topology design for cross-silo federated learning using the theory of max-plus linear systems to compute the system throughput---number of communication rounds per time unit. We also propose practical algorithms that, under the knowledge of measurable network characteristics, find a topology with the largest throughput or with provable throughput guarantees. In realistic Internet networks with 10~Gbps access links for silos, our algorithms speed up training by a factor 9 and 1.5 in comparison to the master-slave architecture and to state-of-the-art MATCHA, respectively. Speedups are even larger with slower access links.

[Quantized Variational Inference](#)

- Amir Dib
- abstract@[open-review](#): We present Quantized Variational Inference, a new algorithm for Evidence Lower Bound minimization. We show how Optimal Voronoi Tesselation produces variance free gradients for Evidence Lower Bound (ELBO) optimization at the cost of introducing asymptotically decaying bias. Subsequently, we propose a Richardson extrapolation type method to improve this bound. We show that using the Quantized Variational Inference framework leads to fast convergence for both score function and the reparametrized gradient estimator at a comparable computational cost. Finally, we propose several experiments to assess the performance of our method and its limitations.

[Asymptotically Optimal Exact Minibatch Metropolis-Hastings](#)

- Ruqi Zhang, A. Feder Cooper, Christopher M. De Sa
- abstract@[open-review](#): Metropolis-Hastings (MH) is a commonly-used MCMC algorithm, but it can be intractable on large datasets due to requiring computations over the whole dataset. In this paper, we study \emph{minibatch MH} methods, which instead use subsamples to enable scaling. We observe that most existing minibatch MH methods are inexact (i.e. they may change the target distribution), and show that this inexactness can cause arbitrarily large errors in inference. We propose a new exact minibatch MH method, \emph{TunaMH}, which exposes a tunable trade-off between its minibatch size and its theoretically guaranteed convergence rate. We prove a lower bound on the batch size that any minibatch MH method \emph{must} use to retain exactness while guaranteeing fast convergence---the first such bound for minibatch MH---and show TunaMH is asymptotically optimal in terms of the batch size. Empirically, we show TunaMH outperforms other exact minibatch MH methods on robust linear regression, truncated Gaussian mixtures, and logistic regression.

[Learning Search Space Partition for Black-box Optimization using Monte Carlo Tree Search](#)

- Linnan Wang, Rodrigo Fonseca, Yuandong Tian
- abstract@[open-review](#): High dimensional black-box optimization has broad applications but remains a challenging problem to solve. Given a set of samples x_i, y_i , building a global model (like Bayesian Optimization (BO)) suffers from the curse of dimensionality in the high-dimensional search space, while a greedy search may lead to sub-optimality. By recursively splitting the search space into regions with high/low function values, recent works like LaNAS shows good performance in Neural Architecture Search (NAS), reducing the sample complexity empirically. In this paper, we coin LA-MCTS that extends LaNAS to other domains. Unlike previous approaches, LA-MCTS learns the partition of the search space using a few samples and their function values in an online fashion. While LaNAS uses linear partition and performs uniform sampling in each region, our LA-MCTS adopts a nonlinear decision boundary and learns a local model to pick good candidates. If the nonlinear partition function and the local model fits well with ground-truth black-box function, then good partitions and candidates can be reached with much fewer samples. LA-MCTS serves as a meta-algorithm by using existing black-box optimizers (e.g., BO, TuRBO as its local models, achieving strong performance in general black-box optimization and reinforcement learning benchmarks, in particular for high-dimensional problems.

[Feature Shift Detection: Localizing Which Features Have Shifted via Conditional Distribution Tests](#)

- Sean Kulinski, Saurabh Bagchi, David I. Inouye
- abstract@[open-review](#): While previous distribution shift detection approaches can identify if a shift has occurred, these approaches cannot localize which specific features have caused a distribution shift---a critical step in diagnosing or fixing any underlying issue. For example, in military sensor networks, users will want to detect when one or more of the sensors has been compromised, and critically, they will want to know which specific sensors might be compromised. Thus, we first define a formalization of this problem as multiple conditional distribution hypothesis tests and propose both non-parametric and parametric statistical tests. For both efficiency and flexibility, we then propose to use a test statistic based on the density model score function (\ie gradient with respect to the input)---which can easily compute test statistics for all dimensions in a single forward and backward pass. Any density model could be used for computing the necessary statistics including deep density models such as normalizing flows or autoregressive models. We additionally develop methods for identifying when and where a shift occurs in multivariate time-series data and show results for multiple scenarios using realistic attack models on both simulated and real-world data.

[Unifying Activation- and Timing-based Learning Rules for Spiking Neural Networks](#)

- Jinseok Kim, Kyungsu Kim, Jae-Joon Kim
- abstract@[open-review](#): For the gradient computation across the time domain in Spiking Neural Networks (SNNs) training, two different approaches have been independently studied. The first is to compute the gradients with respect to the change in spike activation (activation-based methods), and the second is to compute the gradients with respect to the change in spike timing (timing-based methods). In this work, we present a comparative study of the two methods and propose a new supervised learning method that combines them. The proposed method utilizes each individual spike more effectively by shifting spike timings as in the timing-based methods as well as generating and removing spikes as in the activation-based methods. Experimental results showed that the proposed method achieves higher performance in terms of both accuracy and efficiency than the previous approaches.

[Space-Time Correspondence as a Contrastive Random Walk](#)

- Allan Jabri, Andrew Owens, Alexei Efros
- abstract@[open-review](#): This paper proposes a simple self-supervised approach for learning a representation for visual correspondence from raw video. We cast correspondence as prediction of links in a space-time graph constructed from video. In this graph, the nodes are patches sampled from each frame, and nodes adjacent in time can share a directed edge. We learn a representation in which pairwise similarity defines transition probability of a random walk, such that prediction of long-range correspondence is computed as a walk along the graph. We optimize the representation to place high probability along paths of similarity. Targets for learning are formed without supervision, by cycle-consistency: the objective is to maximize the likelihood of returning to the initial node when walking along a graph constructed from a palindrome of frames. Thus, a single path-level constraint implicitly supervises chains of intermediate comparisons. When used as a similarity metric without adaptation, the learned representation outperforms the self-supervised state-of-the-art on label propagation tasks involving objects, semantic parts, and pose. Moreover, we demonstrate that a technique we call edge dropout, as well as self-supervised adaptation at test-time, further improve transfer for object-centric correspondence.

[The Flajolet-Martin Sketch Itself Preserves Differential Privacy: Private Counting with Minimal Space](#)

- Adam Smith, Shuang Song, Abhradeep Guha Thakurta
- abstract@[open-review](#): We revisit the problem of counting the number of distinct elements dist in a data stream D , over a domain $[u]$. We propose an (ϵ, δ) -differentially private algorithm that approximates dist within a factor of $(1/\rho\gamma)^2$, and with additive error of $O(\sqrt{\ln(1/\delta)} / \epsilon)$, using space $O(\ln(\ln(u)/\gamma) / \gamma^2)$. We improve on the prior work at least quadratically and up to exponentially, in terms of both space and additive error. Our additive error guarantee is optimal up to a factor of $O(\sqrt{\ln(1/\delta)})$, and the space bound is optimal up to a factor of $O(\min(\ln(\frac{1}{\delta}) / \gamma, \frac{1}{\gamma^2}))$. We assume the existence of an ideal uniform random hash function, and ignore the space required to store it. We later relax this requirement by assuming pseudorandom functions and appealing to a computational variant of differential privacy, SIM-CDP.

Our algorithm is built on top of the celebrated Flajolet-Martin (FM) sketch. We show that FM-sketch is differentially private as is, as long as there are $\approx \sqrt{\ln(1/\delta)} / \epsilon$ distinct elements in the data set. Along the way, we prove a structural result showing that the maximum of k i.i.d. random variables is statistically close (in the sense of ϵ -differential privacy) to the maximum of $(k+1)$ i.i.d. samples from the same distribution, as long as $k = O(\sqrt{\ln(1/\delta)} / \epsilon)$.

Finally, experiments show that our algorithms introduces error within an order of magnitude of the non-private analogues for streams with thousands of distinct elements, even while providing strong privacy guarantee ($\epsilon \leq 1$).

[Exponential ergodicity of mirror-Langevin diffusions](#)

- Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, Philippe Rigollet, Austin Stromme
- abstract@[open-review](#): Motivated by the problem of sampling from ill-conditioned log-concave distributions, we give a clean non-asymptotic convergence analysis of mirror-Langevin diffusions as introduced in Zhang et al. (2020). As a special case of this framework, we propose a class of diffusions called Newton-Langevin diffusions and prove that they converge to stationarity exponentially fast with a rate which not only is dimension-free, but also has no dependence on the target distribution. We give an application of this result to the problem of sampling from the uniform distribution on a convex body using a strategy inspired by interior-point methods. Our general approach follows the recent trend of linking sampling and optimization and highlights the role of the chi-squared divergence. In particular, it yields new results on the convergence of the vanilla Langevin diffusion in Wasserstein distance.

[An Efficient Framework for Clustered Federated Learning](#)

- Avishek Ghosh, Jichan Chung, Dong Yin, Kannan Ramchandran
- abstract@[open-review](#): We address the problem of Federated Learning (FL) where users are distributed and partitioned into clusters. This setup captures settings where different groups of users have their own objectives (learning tasks) but by aggregating their data with others in the same cluster (same learning task), they can leverage the strength in numbers in order to perform more efficient Federated Learning. We propose a new framework dubbed the Iterative Federated Clustering Algorithm (IFCA), which alternately estimates the cluster identities of the users and optimizes model parameters for the user clusters via gradient descent. We analyze the convergence rate of this algorithm first in a linear model with squared loss and then for generic strongly convex and smooth loss functions. We show that in both settings, with good initialization, IFCA converges at an exponential rate, and discuss the optimality of the statistical error rate. When the clustering structure is ambiguous, we propose to train the models by combining IFCA with the weight sharing technique in multi-task learning. In the experiments, we show that our algorithm can succeed even if we relax the requirements on initialization with random initialization and multiple restarts. We also present experimental results showing that our algorithm is efficient in non-convex problems such as neural networks. We demonstrate the benefits of IFCA over the baselines on several clustered FL benchmarks.

[Autoencoders that don't overfit towards the Identity](#)

- Harald Steck
- abstract@[open-review](#): Autoencoders (AE) aim to reproduce the output from the input. They may hence tend to overfit towards learning the identity-function between the input and output, i.e., they may predict each feature in the output from itself in the input. This is not useful, however, when AEs are used for prediction tasks in the presence of noise in the data. It may seem intuitively evident that this kind of overfitting is prevented by training a denoising AE, as the dropped-out features have to be predicted from the other features. In this paper, we consider linear autoencoders, as they facilitate analytic solutions, and first show that denoising / dropout actually prevents the overfitting towards the identity-function only to the degree that it is penalized by the induced L2-norm regularization. In the main theorem of this paper, we show that the emphasized denoising AE is indeed capable of completely eliminating the overfitting towards the identity-function. Our derivations reveal several new insights, including the closed-form solution of the full-rank model, as well as a new (near-)orthogonality constraint in the low-rank model. While this constraint is conceptually very different from the regularizers recently proposed, their resulting effects on the learned embeddings are empirically similar. Our experiments on three well-known data-sets corroborate the various theoretical insights derived in this paper.

[Polynomial-Time Computation of Optimal Correlated Equilibria in Two-Player Extensive-Form Games with Public Chance Moves and Beyond](#)

- Gabriele Farina, Tuomas Sandholm
- abstract@[open-review](#): Unlike normal-form games, where correlated equilibria have been studied for more than 45 years, extensive-form correlation is still generally not well understood. Part of the reason for this gap is that the sequential nature of extensive-form games allows for a richness of behaviors and incentives that are not possible in normal-form settings. This richness translates to a significantly different complexity landscape surrounding extensive-form correlated equilibria. As of today, it is known that finding an optimal extensive-form correlated equilibrium (EFCE), extensive-form coarse correlated equilibrium (EFCCE), or normal-form coarse correlated equilibrium (NFCCE) in a two-player extensive-form game is computationally tractable when the game does not include chance moves, and intractable when the game involves chance moves. In this paper we significantly refine this complexity threshold by showing that, in two-player games, an optimal correlated equilibrium can be computed in polynomial time, provided that a certain condition is satisfied. We show that the condition holds, for example, when all chance moves are public, that is, both players observe all chance moves. This implies that an optimal EFCE, EFCCE and NFCCE can be computed in polynomial time in the game size in two-player games with public chance moves.

[Parameterized Explainer for Graph Neural Network](#)

- Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, Xiang Zhang
- abstract@[open-review](#): Despite recent progress in Graph Neural Networks (GNNs), explaining predictions made by GNNs remains a challenging open problem. The leading method mainly addresses the local explanations (i.e., important subgraph structure and node features) to interpret why a GNN model makes the prediction for a single instance, e.g. a node or a graph. As a result, the explanation generated is painstakingly customized for each instance. The unique explanation interpreting each instance independently is not sufficient to provide a global understanding of the learned GNN model, leading to the lack of generalizability and hindering it from being used in the inductive setting. Besides, as it is designed for explaining a single instance, it is challenging to explain a set of instances naturally (e.g., graphs of a given class). In this study, we address these key challenges and propose PGExplainer, a parameterized explainer for GNNs. PGExplainer adopts a deep neural network to parameterize the generation process of explanations, which enables PGExplainer a natural approach to multi-instance explanations. Compared to the existing work, PGExplainer has a better generalization power and can be

utilized in an inductive setting easily. Experiments on both synthetic and real-life datasets show highly competitive performance with up to 24.7% relative improvement in AUC on explaining graph classification over the leading baseline.

[Recursive Inference for Variational Autoencoders](#)

- Minyoung Kim, Vladimir Pavlovic
- abstract@[open-review](#): Inference networks of traditional Variational Autoencoders (VAEs) are typically amortized, resulting in relatively inaccurate posterior approximation compared to instance-wise variational optimization. Recent semi-amortized approaches were proposed to address this drawback; however, their iterative gradient update procedures can be computationally demanding. In this paper, we consider a different approach of building a mixture inference model. We propose a novel recursive mixture estimation algorithm for VAEs that iteratively augments the current mixture with new components so as to maximally reduce the divergence between the variational and the true posteriors. Using the functional gradient approach, we devise an intuitive learning criteria for selecting a new mixture component: the new component has to improve the data likelihood (lower bound) and, at the same time, be as divergent from the current mixture distribution as possible, thus increasing representational diversity. Although there have been similar approaches recently, termed boosted variational inference (BVI), our methods differ from BVI in several aspects, most notably that ours deal with recursive inference in VAEs in the form of amortized inference, while BVI is developed within the standard VI framework, leading to a non-amortized single optimization instance, inappropriate for VAEs. A crucial benefit of our approach is that the inference at test time needs a single feed-forward pass through the mixture inference network, making it significantly faster than the semi-amortized approaches. We show that our approach yields higher test data likelihood than the state-of-the-arts on several benchmark datasets.

[Flexible mean field variational inference using mixtures of non-overlapping exponential families](#)

- Jeffrey Spence
- abstract@[open-review](#): Sparse models are desirable for many applications across diverse domains as they can perform automatic variable selection, aid interpretability, and provide regularization. When fitting sparse models in a Bayesian framework, however, analytically obtaining a posterior distribution over the parameters of interest is intractable for all but the simplest cases. As a result practitioners must rely on either sampling algorithms such as Markov chain Monte Carlo or variational methods to obtain an approximate posterior. Mean field variational inference is a particularly simple and popular framework that is often amenable to analytically deriving closed-form parameter updates. When all distributions in the model are members of exponential families and are conditionally conjugate, optimization schemes can often be derived by hand. Yet, I show that using standard mean field variational inference can fail to produce sensible results for models with sparsity-inducing priors, such as the spike-and-slab. Fortunately, such pathological behavior can be remedied as I show that mixtures of exponential family distributions with non-overlapping support form an exponential family. In particular, any mixture of an exponential family of diffuse distributions and a point mass at zero to model sparsity forms an exponential family. Furthermore, specific choices of these distributions maintain conditional conjugacy. I use two applications to motivate these results: one from statistical genetics that has connections to generalized least squares with a spike-and-slab prior on the regression coefficients; and sparse probabilistic principal component analysis. The theoretical results presented here are broadly applicable beyond these two examples.

[HYDRA: Pruning Adversarially Robust Neural Networks](#)

- Vikash Sehwag, Shiqi Wang, Prateek Mittal, Suman Jana
- abstract@[open-review](#): In safety-critical but computationally resource-constrained applications, deep learning faces two key challenges: lack of robustness against adversarial attacks and large neural network size (often millions of parameters). While the research community has extensively explored the use of robust training and network pruning \emph{independently} to address one of these challenges, only a few recent works have studied them jointly. However, these works inherit a heuristic pruning strategy that was developed for benign training, which performs poorly when integrated with robust training techniques, including adversarial training and verifiable robust training. To overcome this challenge, we propose to make pruning techniques aware of the robust training objective and let the training objective guide the search for which connections to prune. We realize this insight by formulating the pruning objective as an empirical risk minimization problem which is solved efficiently using SGD. We demonstrate that our approach, titled HYDRA, achieves compressed networks with \textit{state-of-the-art} benign and robust accuracy, \textit{simultaneously}. We demonstrate the success of our approach across CIFAR-10, SVHN, and ImageNet dataset with four robust training techniques: iterative adversarial training, randomized smoothing, MixTrain, and CROWN-IBP. We also demonstrate the existence of highly robust sub-networks within non-robust networks.

[NVAE: A Deep Hierarchical Variational Autoencoder](#)

- Arash Vahdat, Jan Kautz
- abstract@[open-review](#): Normalizing flows, autoregressive models, variational autoencoders (VAEs), and deep energy-based models are among competing likelihood-based frameworks for deep generative learning. Among them, VAEs have the advantage of fast and tractable sampling and easy-to-access encoding networks. However, they are currently outperformed by other models such as normalizing flows and autoregressive models. While the majority of the research in VAEs is focused on the statistical challenges, we explore the orthogonal direction of carefully designing neural architectures for hierarchical VAEs. We propose Nouveau VAE (NVAE), a deep hierarchical VAE built for image generation using depth-wise separable convolutions and batch normalization. NVAE is equipped with a residual parameterization of Normal distributions and its training is stabilized by spectral regularization. We show that NVAE achieves state-of-the-art results among non-autoregressive likelihood-based models on the MNIST, CIFAR-10, CelebA 64, and CelebA HQ datasets and it provides a strong baseline on FFHQ. For example, on CIFAR-10, NVAE pushes the state-of-the-art from 2.98 to 2.91 bits per dimension, and it produces high-quality images on CelebA HQ. To the best of our knowledge, NVAE is the first successful VAE applied to natural images as large as 256x256 pixels. The source code is publicly available.

[Can Temporal-Difference and Q-Learning Learn Representation? A Mean-Field Theory](#)

- Yufeng Zhang, Qi Cai, Zhuoran Yang, Yongxin Chen, Zhaoran Wang
- abstract@[open-review](#): Temporal-difference and Q-learning play a key role in deep reinforcement learning, where they are empowered by expressive nonlinear function approximators such as neural networks. At the core of their empirical successes is the learned feature representation, which embeds rich observations, e.g., images and texts, into the latent space that encodes semantic structures. Meanwhile, the evolution of such a feature representation is crucial to the convergence of temporal-difference and Q-learning. In particular, temporal-difference learning converges when the function approximator is linear in a feature representation, which is fixed throughout learning, and possibly diverges otherwise. We aim to answer the following questions: When the function approximator is a neural network, how does the associated feature representation evolve? If it converges, does it converge to the optimal one? We prove that utilizing an overparameterized two-layer neural network, temporal-difference and Q-learning globally minimize the mean-squared projected Bellman error at a sublinear rate. Moreover, the associated feature representation converges to the optimal one, generalizing the previous analysis of Cai et al. (2019) in the neural tangent kernel regime, where the associated feature representation stabilizes at the initial one. The key to our analysis is a mean-field perspective, which connects the evolution of a finite-dimensional parameter to its limiting counterpart over an infinite-dimensional Wasserstein space. Our analysis generalizes to soft Q-learning, which is further connected to policy gradient.

[What Do Neural Networks Learn When Trained With Random Labels?](#)

- Hartmut Maennel, Ibrahim M. Alabdulmohsin, Ilya O. Tolstikhin, Robert Baldock, Olivier Bousquet, Sylvain Gelly, Daniel Keysers

- abstract@[open-review](#): We study deep neural networks (DNNs) trained on natural image data with entirely random labels. Despite its popularity in the literature, where it is often used to study memorization, generalization, and other phenomena, little is known about what DNNs learn in this setting. In this paper, we show analytically for convolutional and fully connected networks that an alignment between the principal components of network parameters and data takes place when training with random labels. We study this alignment effect by investigating neural networks pre-trained on randomly labelled image data and subsequently fine-tuned on disjoint datasets with random or real labels. We show how this alignment produces a positive transfer: networks pre-trained with random labels train faster downstream compared to training from scratch even after accounting for simple effects, such as weight scaling. We analyze how competing effects, such as specialization at later layers, may hide the positive transfer. These effects are studied in several network architectures, including VGG16 and ResNet18, on CIFAR10 and ImageNet.

[Counterfactual Prediction for Bundle Treatment](#)

- Hao Zou, Peng Cui, Bo Li, Zheyuan Shen, Jianxin Ma, Hongxia Yang, Yue He
- abstract@[open-review](#): Estimating counterfactual outcome of different treatments from observational data is an important problem to assist decision making in a variety of fields. Among the various forms of treatment specification, bundle treatment has been widely adopted in many scenarios, such as recommendation systems and online marketing. The bundle treatment usually can be abstracted as a high dimensional binary vector, which makes it more challenging for researchers to remove the confounding bias in observational data. In this work, we assume the existence of low dimensional latent structure underlying bundle treatment. Via the learned latent representations of treatments, we propose a novel variational sample re-weighting (VSR) method to eliminate confounding bias by decorrelating the treatments and confounders. Finally, we conduct extensive experiments to demonstrate that the predictive model trained on this re-weighted dataset can achieve more accurate counterfactual outcome prediction.

[Beta Embeddings for Multi-Hop Logical Reasoning in Knowledge Graphs](#)

- Hongyu Ren, Jure Leskovec
- abstract@[open-review](#): One of the fundamental problems in Artificial Intelligence is to perform complex multi-hop logical reasoning over the facts captured by a knowledge graph (KG). This problem is challenging, because KGs can be massive and incomplete. Recent approaches embed KG entities in a low dimensional space and then use these embeddings to find the answer entities. However, it has been an outstanding challenge of how to handle arbitrary first-order logic (FOL) queries as present methods are limited to only a subset of FOL operators. In particular, the negation operator is not supported. An additional limitation of present methods is also that they cannot naturally model uncertainty. Here, we present BetaE, a probabilistic embedding framework for answering arbitrary FOL queries over KGs. BetaE is the first method that can handle a complete set of first-order logical operations: conjunction (\wedge), disjunction (\vee), and negation (\neg). A key insight of BetaE is to use probabilistic distributions with bounded support, specifically the Beta distribution, and embed queries/entities as distributions, which as a consequence allows us to also faithfully model uncertainty. Logical operations are performed in the embedding space by neural operators over the probabilistic embeddings. We demonstrate the performance of BetaE on answering arbitrary FOL queries on three large, incomplete KGs. While being more general, BetaE also increases relative performance by up to 25.4% over the current state-of-the-art KG reasoning methods that can only handle conjunctive queries without negation.

[Learning Disentangled Representations and Group Structure of Dynamical Environments](#)

- Robin Quessard, Thomas Barrett, William Clements
- abstract@[open-review](#): Learning disentangled representations is a key step towards effectively discovering and modelling the underlying structure of environments. In the natural sciences, physics has found great success by describing the universe in terms of symmetry preserving transformations. Inspired by this formalism, we propose a framework, built upon the theory of group representation, for learning representations of a dynamical environment structured around the transformations that generate its evolution. Experimentally, we learn the structure of explicitly symmetric environments without supervision from observational data generated by sequential interactions. We further introduce an intuitive disentanglement regularisation to ensure the interpretability of the learnt representations. We show that our method enables accurate long-horizon predictions, and demonstrate a correlation between the quality of predictions and disentanglement in the latent space.

[Learning Linear Programs from Optimal Decisions](#)

- Yingcong Tan, Daria Terekhov, Andrew Delong
- abstract@[open-review](#): We propose a flexible gradient-based framework for learning linear programs from optimal decisions. Linear programs are often specified by hand, using prior knowledge of relevant costs and constraints. In some applications, linear programs must instead be learned from observations of optimal decisions. Learning from optimal decisions is a particularly challenging bilevel problem, and much of the related inverse optimization literature is dedicated to special cases. We tackle the general problem, learning all parameters jointly while allowing flexible parameterizations of costs, constraints, and loss functions. We also address challenges specific to learning linear programs, such as empty feasible regions and non-unique optimal decisions. Experiments show that our method successfully learns synthetic linear programs and minimum-cost multi-commodity flow instances for which previous methods are not directly applicable. We also provide a fast batch-mode PyTorch implementation of the homogeneous interior point algorithm, which supports gradients by implicit differentiation or backpropagation.

[Wisdom of the Ensemble: Improving Consistency of Deep Learning Models](#)

- Lijing Wang, Dipanjan Ghosh, Maria Gonzalez Diaz, Ahmed Farahat, Mahbubul Alam, Chetan Gupta, Jiangzhuo Chen, Madhav Marathe
- abstract@[open-review](#): Deep learning classifiers are assisting humans in making decisions and hence the user's trust in these models is of paramount importance. Trust is often a function of constant behavior. From an AI model perspective it means given the same input the user would expect the same output, especially for correct outputs, or in other words consistently correct outputs. This paper studies a model behavior in the context of periodic retraining of deployed models where the outputs from successive generations of the models might not agree on the correct labels assigned to the same input. We formally define consistency and correct-consistency of a learning model. We prove that consistency and correct-consistency of an ensemble learner is not less than the average consistency and correct-consistency of individual learners and correct-consistency can be improved with a probability by combining learners with accuracy not less than the average accuracy of ensemble component learners. To validate the theory using three datasets and two state-of-the-art deep learning classifiers we also propose an efficient dynamic snapshot ensemble method and demonstrate its value. Code for our algorithm is available at <https://github.com/christa60/dynens>.

[Universal Function Approximation on Graphs](#)

- Rickard Bräiel Gabrielsson
- abstract@[open-review](#): In this work we produce a framework for constructing universal function approximators on graph isomorphism classes. We prove how this framework comes with a collection of theoretically desirable properties and enables novel analysis. We show how this allows us to achieve state-of-the-art performance on four different well-known datasets in graph classification and separate classes of graphs that other graph-learning methods cannot. Our approach is inspired by persistent homology, dependency parsing for NLP, and multivalued functions. The complexity of the underlying algorithm is $O(\#edges \times \#nodes)$ and code is publicly available (<https://github.com/bruel-gabrielsson/universal-function-approximation-on-graphs>).

[Accelerating Reinforcement Learning through GPU Atari Emulation](#)

- Steven Dalton, iuri frosio
- abstract@[open-review](#): We introduce CuLE (CUDA Learning Environment), a CUDA port of the Atari Learning Environment (ALE) which is used for the development of deep reinforcement algorithms. CuLE overcomes many limitations of existing CPU-based emulators and scales naturally to multiple GPUs. It leverages GPU parallelization to run thousands of games simultaneously and it renders frames directly on the GPU, to avoid the bottleneck arising from the limited CPU-GPU communication bandwidth. CuLE generates up to 155M frames per hour on a single GPU, a finding previously achieved only through a cluster of CPUs. Beyond highlighting the differences between CPU and GPU emulators in the context of reinforcement learning, we show how to leverage the high throughput of CuLE by effective batching of the training data, and show accelerated convergence for A2C+V-trace. CuLE is available at <https://github.com/NVlabs/cule>.

[EvolveGraph: Multi-Agent Trajectory Prediction with Dynamic Relational Reasoning](#)

- Jiachen Li, Fan Yang, Masayoshi Tomizuka, Chiho Choi
- abstract@[open-review](#): Multi-agent interacting systems are prevalent in the world, from purely physical systems to complicated social dynamic systems. In many applications, effective understanding of the situation and accurate trajectory prediction of interactive agents play a significant role in downstream tasks, such as decision making and planning. In this paper, we propose a generic trajectory forecasting framework (named EvolveGraph) with explicit relational structure recognition and prediction via latent interaction graphs among multiple heterogeneous, interactive agents. Considering the uncertainty of future behaviors, the model is designed to provide multi-modal prediction hypotheses. Since the underlying interactions may evolve even with abrupt changes, and different modalities of evolution may lead to different outcomes, we address the necessity of dynamic relational reasoning and adaptively evolving the interaction graphs. We also introduce a double-stage training pipeline which not only improves training efficiency and accelerates convergence, but also enhances model performance. The proposed framework is evaluated on both synthetic physics simulations and multiple real-world benchmark datasets in various areas. The experimental results illustrate that our approach achieves state-of-the-art performance in terms of prediction accuracy.

[Comparator-Adaptive Convex Bandits](#)

- Dirk van der Hoeven, Ashok Cutkosky, Haipeng Luo
- abstract@[open-review](#): We study bandit convex optimization methods that adapt to the norm of the comparator, a topic that has only been studied before for its full-information counterpart. Specifically, we develop convex bandit algorithms with regret bounds that are small whenever the norm of the comparator is small. We first use techniques from the full-information setting to develop comparator-adaptive algorithms for linear bandits. Then, we extend the ideas to convex bandits with Lipschitz or smooth loss functions, using a new single-point gradient estimator and carefully designed surrogate losses.

[Model-based Reinforcement Learning for Semi-Markov Decision Processes with Neural ODEs](#)

- Jianzhun Du, Joseph Futoma, Finale Doshi-Velez
- abstract@[open-review](#): We present two elegant solutions for modeling continuous-time dynamics, in a novel model-based reinforcement learning (RL) framework for semi-Markov decision processes (SMDPs), using neural ordinary differential equations (ODEs). Our models accurately characterize continuous-time dynamics and enable us to develop high-performing policies using a small amount of data. We also develop a model-based approach for optimizing time schedules to reduce interaction rates with the environment while maintaining the near-optimal performance, which is not possible for model-free methods. We experimentally demonstrate the efficacy of our methods across various continuous-time domains.

[The Adaptive Complexity of Maximizing a Gross Substitutes Valuation](#)

- Ron Kupfer, Sharon Qian, Eric Balkanski, Yaron Singer
- abstract@[open-review](#): In this paper, we study the adaptive complexity of maximizing a monotone gross substitutes function under a cardinality constraint. Our main result is an algorithm that achieves a 1-epsilon approximation in O(log n) adaptive rounds for any constant epsilon > 0, which is an exponential speedup in parallel running time compared to previously studied algorithms for gross substitutes functions. We show that the algorithmic results are tight in the sense that there is no algorithm that obtains a constant factor approximation in o(log n) rounds. Both the upper and lower bounds are under the assumption that queries are only on feasible sets (i.e., of size at most k). We also show that under a stronger model, where non-feasible queries are allowed, there is no non-adaptive algorithm that obtains an approximation better than 1/2 + epsilon. Both lower bounds extend to the class of OXS functions. Additionally, we conduct experiments on synthetic and real data sets to demonstrate the near-optimal performance and efficiency of the algorithm in practice.

[A Robust Functional EM Algorithm for Incomplete Panel Count Data](#)

- Alexander Moreno, Zhenke Wu, Jamie Roslyn Yap, Cho Lam, David Wetter, Inbal Nahum-Shani, Walter Dempsey, James M. Rehg
- abstract@[open-review](#): Panel count data describes aggregated counts of recurrent events observed at discrete time points. To understand dynamics of health behaviors and predict future negative events, the field of quantitative behavioral research has evolved to increasingly rely upon panel count data collected via multiple self reports, for example, about frequencies of smoking using in-the-moment surveys on mobile devices. However, missing reports are common and present a major barrier to downstream statistical learning. As a first step, under a missing completely at random assumption (MCAR), we propose a simple yet widely applicable functional EM algorithm to estimate the counting process mean function, which is of central interest to behavioral scientists. The proposed approach wraps several popular panel count inference methods, seamlessly deals with incomplete counts and is robust to misspecification of the Poisson process assumption. Theoretical analysis of the proposed algorithm provides finite-sample guarantees by extending parametric EM theory to the general non-parametric setting. We illustrate the utility of the proposed algorithm through numerical experiments and an analysis of smoking cessation data. We also discuss useful extensions to address deviations from the MCAR assumption and covariate effects.

[Graph Stochastic Neural Networks for Semi-supervised Learning](#)

- Haibo Wang, Chuan Zhou, Xin Chen, Jia Wu, Shirui Pan, Jilong Wang
- abstract@[open-review](#): Graph Neural Networks (GNNs) have achieved remarkable performance in the task of the semi-supervised node classification. However, most existing models learn a deterministic classification function, which lack sufficient flexibility to explore better choices in the presence of kinds of imperfect observed data such as the scarce labeled nodes and noisy graph structure. To improve the rigidness and inflexibility of deterministic classification functions, this paper proposes a novel framework named Graph Stochastic Neural Networks (GSNN), which aims to model the uncertainty of the classification function by simultaneously learning a family of functions, i.e., a stochastic function. Specifically, we introduce a learnable graph neural network coupled with a high-dimensional latent variable to model the distribution of the classification function, and further adopt the amortised variational inference to approximate the intractable joint posterior for missing labels and the latent variable. By maximizing the lower-bound of the likelihood for observed node labels, the instantiated models can be trained in an end-to-end manner effectively. Extensive experiments on three real-world datasets show that GSNN achieves substantial performance gain in different scenarios compared with stat-of-the-art baselines.

[Compositional Zero-Shot Learning via Fine-Grained Dense Feature Composition](#)

- Dat Huynh, Ehsan Elhamifar
- abstract@[open-review](#): We develop a novel generative model for zero-shot learning to recognize fine-grained unseen classes without training samples. Our observation is that generating holistic features of unseen classes fails to capture every attribute needed to distinguish small differences among classes. We propose a feature composition framework that learns to extract attribute-based features from training samples and combines them to construct fine-grained features for unseen classes. Feature composition allows us to not only selectively compose features of unseen classes from only relevant training samples, but also obtain diversity among composed features via changing samples used for composition. In addition, instead of building a global feature of an unseen class, we use all attribute-based features to form a dense representation consisting of fine-grained attribute details. To recognize unseen classes, we propose a novel training scheme that uses a discriminative model to construct features that are subsequently used to train itself. Therefore, we directly train the discriminative model on composed features without learning separate generative models. We conduct experiments on four popular datasets of DeepFashion, AWA2, CUB, and SUN, showing that our method significantly improves the state of the art.

[A Benchmark for Systematic Generalization in Grounded Language Understanding](#)

- Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, Brenden M. Lake
- abstract@[open-review](#): Humans easily interpret expressions that describe unfamiliar situations composed from familiar parts ("greet the pink brontosaurus by the ferris wheel"). Modern neural networks, by contrast, struggle to interpret novel compositions. In this paper, we introduce a new benchmark, gSCAN, for evaluating compositional generalization in situated language understanding. Going beyond a related benchmark that focused on syntactic aspects of generalization, gSCAN defines a language grounded in the states of a grid world, facilitating novel evaluations of acquiring linguistically motivated rules. For example, agents must understand how adjectives such as 'small' are interpreted relative to the current world state or how adverbs such as 'cautiously' combine with new verbs. We test a strong multi-modal baseline model and a state-of-the-art compositional method finding that, in most cases, they fail dramatically when generalization requires systematic compositional rules.

[Weston-Watkins Hinge Loss and Ordered Partitions](#)

- Yutong Wang, Clayton Scott
- abstract@[open-review](#): Multiclass extensions of the support vector machine (SVM) have been formulated in a variety of ways. A recent empirical comparison of nine such formulations [DoC̄san et al. 2016] recommends the variant proposed by Weston and Watkins (WW), despite the fact that the WW-hinge loss is not calibrated with respect to the 0-1 loss. In this work we introduce a novel discrete loss function for multiclass classification, the ordered partition loss, and prove that the WW-hinge loss is calibrated with respect to this loss. We also argue that the ordered partition loss is minimally emblematic among discrete losses satisfying this property. Finally, we apply our theory to justify the empirical observation made by DoC̄san et al that the WW-SVM can work well even under massive label noise, a challenging setting for multiclass SVMs.

[Reinforcement Learning with Augmented Data](#)

- Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, Aravind Srinivas
- abstract@[open-review](#): Learning from visual observations is a fundamental yet challenging problem in Reinforcement Learning (RL). Although algorithmic advances combined with convolutional neural networks have proved to be a recipe for success, current methods are still lacking on two fronts: (a) data-efficiency of learning and (b) generalization to new environments. To this end, we present Reinforcement Learning with Augmented Data (RAD), a simple plug-and-play module that can enhance most RL algorithms. We perform the first extensive study of general data augmentations for RL on both pixel-based and state-based inputs, and introduce two new data augmentations - random translate and random amplitude scale. We show that augmentations such as random translate, crop, color jitter, patch cutout, random convolutions, and amplitude scale can enable simple RL algorithms to outperform complex state-of-the-art methods across common benchmarks. RAD sets a new state-of-the-art in terms of data-efficiency and final performance on the DeepMind Control Suite benchmark for pixel-based control as well as OpenAI Gym benchmark for state-based control. We further demonstrate that RAD significantly improves test-time generalization over existing methods on several OpenAI ProcGen benchmarks.

[Towards Minimax Optimal Reinforcement Learning in Factored Markov Decision Processes](#)

- Yi Tian, Jian Qian, Suvrit Sra
- abstract@[open-review](#): We study minimax optimal reinforcement learning in episodic factored Markov decision processes (FMDPs), which are MDPs with conditionally independent transition components. Assuming the factorization is known, we propose two model-based algorithms. The first one achieves minimax optimal regret guarantees for a rich class of factored structures, while the second one enjoys better computational complexity with a slightly worse regret. A key new ingredient of our algorithms is the design of a bonus term to guide exploration. We complement our algorithms by presenting several structure dependent lower bounds on regret for FMDPs that reveal the difficulty hiding in the intricacy of the structures.

[Graduated Assignment for Joint Multi-Graph Matching and Clustering with Application to Unsupervised Graph Matching Network Learning](#)

- Runzhong Wang, Junchi Yan, Xiaokang Yang
- abstract@[open-review](#): This paper considers the setting of jointly matching and clustering multiple graphs belonging to different groups, which naturally rises in many realistic problems. Both graph matching and clustering are challenging (NP-hard) and a joint solution is appealing due to the natural connection of the two tasks. In this paper, we resort to a graduated assignment procedure for soft matching and clustering over iterations, whereby the two-way constraint and clustering confidence are modulated by two separate annealing parameters, respectively. Our technique can be further utilized for end-to-end learning whose loss refers to the cross-entropy between two lines of matching pipelines, as such the keypoint feature extraction CNNs can be learned without ground-truth supervision. Experimental results on real-world benchmarks show our method outperforms learning-free algorithms and performs comparatively against two-graph based supervised graph matching approaches.

[Estimating Training Data Influence by Tracing Gradient Descent](#)

- Garima Pruthi, Frederick Liu, Satyen Kale, Mukund Sundararajan
- abstract@[open-review](#): We introduce a method called TracIn that computes the influence of a training example on a prediction made by the model. The idea is to trace how the loss on the test point changes during the training process whenever the training example of interest was utilized. We provide a scalable implementation of TracIn via: (a) a first-order gradient approximation to the exact computation, (b) saved checkpoints of standard training procedures, and (c) cherry-picking layers of a deep neural network. In contrast with previously proposed methods, TracIn is simple to implement; all it needs is the ability to work with gradients, checkpoints, and loss functions. The method is general. It applies to any machine learning model trained using stochastic gradient descent or a variant of it, agnostic of architecture, domain and task. We expect the method to be widely useful within processes that study and improve training data.

[Joint Policy Search for Multi-agent Collaboration with Imperfect Information](#)

- Yuandong Tian, Qucheng Gong, Yu Jiang

- abstract@[open-review](#): To learn good joint policies for multi-agent collaboration with incomplete information remains a fundamental challenge. While for two-player zero-sum games, coordinate-ascent approaches (optimizing one agent's policy at a time, e.g., self-play) work with guarantees, in multi-agent cooperative setting they often converge to sub-optimal Nash equilibrium. On the other hand, directly modeling joint policy changes in incomplete information game is nontrivial due to complicated interplay of policies (e.g., upstream updates affect downstream state reachability). In this paper, we show global changes of game values can be decomposed to policy changes localized at each information set, with a novel term named \text{policy-change density}. Based on this, we propose \text{Joint Policy Search} (JPS) that iteratively improves joint policies of collaborative agents in incomplete information games, without re-evaluating the entire game. On multiple collaborative tabular games, JPS is proven to never worsen performance and can improve solutions provided by unilateral approaches (e.g. CFR), outperforming algorithms designed for collaborative policy learning (e.g. BAD). Furthermore, for real-world game whose states are too many to enumerate, \text{ours} has an online form that naturally links with gradient updates. We test it to Contract Bridge, a 4-player imperfect-information game where a team of \$2\$ collaborates to compete against the other. In its bidding phase, players bid in turn to find a good contract through a limited information channel. Based on a strong baseline agent that bids competitive bridge purely through domain-agnostic self-play, JPS improves collaboration of team players and outperforms WBridge5, a championship-winning software, by \$+0.63\$ IMPs (International Matching Points) per board over \$1000\$ games, substantially better than previous SoTA (\$+0.41\$ IMPs/b against WBridge5). Note that \$+0.1\$ IMPs/b is regarded as a nontrivial improvement in Computer Bridge.

[Adversarial Bandits with Corruptions: Regret Lower Bound and No-regret Algorithm](#)

- lin yang, Mohammad Hajiesmaili, Mohammad Sadegh Talebi, John C. S. Lui, Wing Shing Wong
- abstract@[open-review](#): This paper studies adversarial bandit setting, the reward of arms is predetermined by an adversary who is oblivious to the learner's policy. In this paper, we consider an extended setting in which an attacker sits in-between the environment and the learner, and is endowed with a limited budget to corrupt the reward of the selected arm. We have two main results. First, we derive a lower bound on the regret of any bandit algorithm that is aware of the budget of the attacker. Also, for budget-agnostic algorithms, we characterize an impossibility result demonstrating that even when the attacker has a sublinear budget, i.e., a budget growing sublinearly with time horizon T, they fail to achieve a sublinear regret. Second, we propose ExpRb, a bandit algorithm that incorporates a biased estimator and a robustness parameter to deal with corruption. We characterize the regret of ExpRb as a function of the corruption budget and show that for the case of a known corruption budget, the regret of ExpRb is tight.

[Beta R-CNN: Looking into Pedestrian Detection from Another Perspective](#)

- Zixuan Xu, Banghuai Li, Ye Yuan, Anhong Dang
- abstract@[open-review](#): Recently significant progress has been made in pedestrian detection, but it remains challenging to achieve high performance in occluded and crowded scenes. It could be mostly attributed to the widely used representation of pedestrians, i.e., 2D axis-aligned bounding box, which just describes the approximate location and size of the object. Bounding box models the object as a uniform distribution within the boundary, making pedestrians indistinguishable in occluded and crowded scenes due to much noise. To eliminate the problem, we propose a novel representation based on 2D beta distribution, named Beta Representation. It pictures a pedestrian by explicitly constructing the relationship between full-body and visible boxes, and emphasizes the center of visual mass by assigning different probability values to pixels. As a result, Beta Representation is much better for distinguishing highly-overlapped instances in crowded scenes with a new NMS strategy named BetaNMS. What's more, to fully exploit Beta Representation, a novel pipeline Beta R-CNN equipped with BetaHead and BetaMask is proposed, leading to high detection performance in occluded and crowded scenes.

[Batch Normalization Biases Residual Blocks Towards the Identity Function in Deep Networks](#)

- Soham De, Sam Smith
- abstract@[open-review](#): Batch normalization dramatically increases the largest trainable depth of residual networks, and this benefit has been crucial to the empirical success of deep residual networks on a wide range of benchmarks. We show that this key benefit arises because, at initialization, batch normalization down-scales the residual branch relative to the skip connection, by a normalizing factor on the order of the square root of the network depth. This ensures that, early in training, the function computed by normalized residual blocks in deep networks is close to the identity function (on average). We use this insight to develop a simple initialization scheme that can train deep residual networks without normalization. We also provide a detailed empirical study of residual networks, which clarifies that, although batch normalized networks can be trained with larger learning rates, this effect is only beneficial in specific compute regimes, and has minimal benefits when the batch size is small.

[Learning Retrospective Knowledge with Reverse Reinforcement Learning](#)

- Shangtong Zhang, Vivek Veeriah, Shimon Whiteson
- abstract@[open-review](#): We present a Reverse Reinforcement Learning (Reverse RL) approach for representing retrospective knowledge. General Value Functions (GVFs) have enjoyed great success in representing predictive knowledge, i.e., answering questions about possible future outcomes such as "how much fuel will be consumed in expectation if we drive from A to B?". GVF, however, cannot answer questions like "how much fuel do we expect a car to have given it is at B at time t?". To answer this question, we need to know when that car had a full tank and how that car came to B. Since such questions emphasize the influence of possible past events on the present, we refer to their answers as retrospective knowledge. In this paper, we show how to represent retrospective knowledge with Reverse GVF, which are trained via Reverse RL. We demonstrate empirically the utility of Reverse GVF in both representation learning and anomaly detection.

[Dialog without Dialog Data: Learning Visual Dialog Agents from VQA Data](#)

- Michael Cogswell, Jiasen Lu, Rishabh Jain, Stefan Lee, Devi Parikh, Dhruv Batra
- abstract@[open-review](#): Can we develop visually grounded dialog agents that can efficiently adapt to new tasks without forgetting how to talk to people? Such agents could leverage a larger variety of existing data to generalize to a new task, minimizing expensive data collection and annotation. In this work, we study a setting we call "Dialog without Dialog", which requires agents to develop visually grounded dialog models that can adapt to new tasks without language level supervision. By factorizing intention and language, our model minimizes linguistic drift after fine-tuning for new tasks. We present qualitative results, automated metrics, and human studies that all show our model can adapt to new tasks and maintain language quality. Baselines either fail to perform well at new tasks or experience language drift, becoming unintelligible to humans. Code has been made available at: <https://github.com/mcogswell/dialogwithoutdialog>.

[GCOMB: Learning Budget-constrained Combinatorial Algorithms over Billion-sized Graphs](#)

- Sahil Manchanda, AKASH MITTAL, Anuj Dhawan, Sourav Medya, Sayan Ranu, Ambuj Singh
- abstract@[open-review](#): There has been an increased interest in discovering heuristics for combinatorial problems on graphs through machine learning. While existing techniques have primarily focused on obtaining high-quality solutions, scalability to billion-sized graphs has not been adequately addressed. In addition, the impact of a budget-constraint, which is necessary for many practical scenarios, remains to be studied. In this paper, we propose a framework called GCOMB to bridge these gaps. GCOMB trains a Graph Convolutional Network (GCN) using a novel probabilistic greedy mechanism to predict the quality of a node. To further facilitate the combinatorial nature of the problem, GCOMB utilizes a Q-learning framework, which is made

efficient through importance sampling. We perform extensive experiments on real graphs to benchmark the efficiency and efficacy of GCOMB. Our results establish that GCOMB is 100 times faster and marginally better in quality than state-of-the-art algorithms for learning combinatorial algorithms. Additionally, a case-study on the practical combinatorial problem of Influence Maximization (IM) shows GCOMB is 150 times faster than the specialized IM algorithm IMM with similar quality.

[A General Large Neighborhood Search Framework for Solving Integer Linear Programs](#)

- Jialin Song, ravi lanka, Yisong Yue, Bistra Dilkina
- abstract@[open-review](#): This paper studies how to design abstractions of large-scale combinatorial optimization problems that can leverage existing state-of-the-art solvers in general-purpose ways, and that are amenable to data-driven design. The goal is to arrive at new approaches that can reliably outperform existing solvers in wall-clock time. We focus on solving integer programs and ground our approach in the large neighborhood search (LNS) paradigm, which iteratively chooses a subset of variables to optimize while leaving the remainder fixed. The appeal of LNS is that it can easily use any existing solver as a subroutine, and thus can inherit the benefits of carefully engineered heuristic approaches and their software implementations. We also show that one can learn a good neighborhood selector from training data. Through an extensive empirical validation, we demonstrate that our LNS framework can significantly outperform, in wall-clock time, compared to state-of-the-art commercial solvers such as Gurobi.

[A Theoretical Framework for Target Propagation](#)

- Alexander Meulemans, Francesco Carzaniga, Johan Suykens, JoÃ£o Sacramento, Benjamin F. Grewe
- abstract@[open-review](#): The success of deep learning, a brain-inspired form of AI, has sparked interest in understanding how the brain could similarly learn across multiple layers of neurons. However, the majority of biologically-plausible learning algorithms have not yet reached the performance of backpropagation (BP), nor are they built on strong theoretical foundations. Here, we analyze target propagation (TP), a popular but not yet fully understood alternative to BP, from the standpoint of mathematical optimization. Our theory shows that TP is closely related to Gauss-Newton optimization and thus substantially differs from BP. Furthermore, our analysis reveals a fundamental limitation of difference target propagation (DTP), a well-known variant of TP, in the realistic scenario of non-invertible neural networks. We provide a first solution to this problem through a novel reconstruction loss that improves feedback weight training, while simultaneously introducing architectural flexibility by allowing for direct feedback connections from the output to each hidden layer. Our theory is corroborated by experimental results that show significant improvements in performance and in the alignment of forward weight updates with loss gradients, compared to DTP.

[OrganITE: Optimal transplant donor organ offering using an individual treatment effect](#)

- Jeroen Berrevoets, James Jordon, Ioana Bica, alexander gimson, Mihaela van der Schaar
- abstract@[open-review](#): Transplant-organs are a scarce medical resource. The uniqueness of each organ and the patients' heterogeneous responses to the organs present a unique and challenging machine learning problem. In this problem there are two key challenges: (i) assigning each organ "optimally" to a patient in the queue; (ii) accurately estimating the potential outcomes associated with each patient and each possible organ. In this paper, we introduce OrganITE, an organ-to-patient assignment methodology that assigns organs based not only on its own estimates of the potential outcomes but also on organ scarcity. By modelling and accounting for organ scarcity we significantly increase total life years across the population, compared to the existing greedy approaches that simply optimise life years for the current organ available. Moreover, we propose an individualised treatment effect model capable of addressing the high dimensionality of the organ space. We test our method on real and simulated data, resulting in as much as an additional year of life expectancy as compared to existing organ-to-patient policies.

[The Complete Lasso Tradeoff Diagram](#)

- Hua Wang, Yachong Yang, Zhiqi Bu, Weijie Su
- abstract@[open-review](#): A fundamental problem in high-dimensional regression is to understand the tradeoff between type I and type II errors or, equivalently, false discovery rate (FDR) and power in variable selection. To address this important problem, we offer the first complete diagram that distinguishes all pairs of FDR and power that can be asymptotically realized by the Lasso from the remaining pairs, in a regime of linear sparsity under random designs. The tradeoff between the FDR and power characterized by our diagram holds no matter how strong the signals are. In particular, our results complete the earlier Lasso tradeoff diagram in previous literature by recognizing two simple constraints on the pairs of FDR and power. The improvement is more substantial when the regression problem is above the Donoho-Tanner phase transition. Finally, we present extensive simulation studies to confirm the sharpness of the complete Lasso tradeoff diagram.

[On the universality of deep learning](#)

- Emmanuel Abbe, Colin Sandon
- abstract@[open-review](#): This paper shows that deep learning, i.e., neural networks trained by SGD, can learn in polytime any function class that can be learned in polytime by some algorithm, including parities. This universal result is further shown to be robust, i.e., it holds under possibly poly-noise on the gradients, which gives a separation between deep learning and statistical query algorithms, as the latter are not comparably universal due to cases like parities. This also shows that SGD-based deep learning does not suffer from the limitations of the perceptron discussed by Minsky-Papert '69. The paper further complement this result with a lower-bound on the generalization error of descent algorithms, which implies in particular that the robust universality breaks down if the gradients are averaged over large enough batches of samples as in full-GD, rather than fewer samples as in SGD.

[Regression with reject option and application to kNN](#)

- Ahmed Zaoui, Christophe Denis, Mohamed Hebiri
- abstract@[open-review](#): We investigate the problem of regression where one is allowed to abstain from predicting. We refer to this framework as regression with reject option as an extension of classification with reject option. In this context, we focus on the case where the rejection rate is fixed and derive the optimal rule which relies on thresholding the conditional variance function. We provide a semi-supervised estimation procedure of the optimal rule involving two datasets: a first labeled dataset is used to estimate both regression function and conditional variance function while a second unlabeled dataset is exploited to calibrate the desired rejection rate. The resulting predictor with reject option is shown to be almost as good as the optimal predictor with reject option both in terms of risk and rejection rate. We additionally apply our methodology with kNN algorithm and establish rates of convergence for the resulting kNN predictor under mild conditions. Finally, a numerical study is performed to illustrate the benefit of using the proposed procedure.

[The Primal-Dual method for Learning Augmented Algorithms](#)

- Etienne Bamas, Andreas Maggiori, Ola Svensson
- abstract@[open-review](#): The extension of classical online algorithms when provided with predictions is a new and active research area. In this paper, we extend the primal-dual method for online algorithms in order to incorporate predictions that advise the online algorithm about the next action to take. We use this framework to obtain novel algorithms for a variety of online covering problems. We compare our algorithms to the cost of the true and predicted offline optimal solutions and show that these algorithms outperform any online algorithm when the prediction is accurate while maintaining good guarantees when the prediction is misleading.

[FLAMBE: Structural Complexity and Representation Learning of Low Rank MDPs](#)

- Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, Wen Sun
- abstract@[open-review](#): In order to deal with the curse of dimensionality in reinforcement learning (RL), it is common practice to make parametric assumptions where values or policies are functions of some low dimensional feature space. This work focuses on the representation learning question: how can we learn such features? Under the assumption that the underlying (unknown) dynamics correspond to a low rank transition matrix, we show how the representation learning question is related to a particular non-linear matrix decomposition problem. Structurally, we make precise connections between these low rank MDPs and latent variable models, showing how they significantly generalize prior formulations, such as block MDPs, for representation learning in RL. Algorithmically, we develop FLAMBE, which engages in exploration and representation learning for provably efficient RL in low rank transition models. On a technical level, our analysis eliminates reachability assumptions that appear in prior results on the simpler block MDP model and may be of independent interest.

[A Class of Algorithms for General Instrumental Variable Models](#)

- Niki Kilbertus, Matt J. Kusner, Ricardo Silva
- abstract@[open-review](#): Causal treatment effect estimation is a key problem that arises in a variety of real-world settings, from personalized medicine to governmental policy making. There has been a flurry of recent work in machine learning on estimating causal effects when one has access to an instrument. However, to achieve identifiability, they in general require one-size-fits-all assumptions such as an additive error model for the outcome. An alternative is partial identification, which provides bounds on the causal effect. Little exists in terms of bounding methods that can deal with the most general case, where the treatment itself can be continuous. Moreover, bounding methods generally do not allow for a continuum of assumptions on the shape of the causal effect that can smoothly trade off stronger background knowledge for more informative bounds. In this work, we provide a method for causal effect bounding in continuous distributions, leveraging recent advances in gradient-based methods for the optimization of computationally intractable objective functions. We demonstrate on a set of synthetic and real-world data that our bounds capture the causal effect when additive methods fail, providing a useful range of answers compatible with observation as opposed to relying on unwarranted structural assumptions.

[Black-Box Ripper: Copying black-box models using generative evolutionary algorithms](#)

- Antonio Barbalau, Adrian Cosma, Radu Tudor Ionescu, Marius Popescu
- abstract@[open-review](#): We study the task of replicating the functionality of black-box neural models, for which we only know the output class probabilities provided for a set of input images. We assume back-propagation through the black-box model is not possible and its training images are not available, e.g. the model could be exposed only through an API. In this context, we present a teacher-student framework that can distill the black-box (teacher) model into a student model with minimal accuracy loss. To generate useful data samples for training the student, our framework (i) learns to generate images on a proxy data set (with images and classes different from those used to train the black-box) and (ii) applies an evolutionary strategy to make sure that each generated data sample exhibits a high response for a specific class when given as input to the black box. Our framework is compared with several baseline and state-of-the-art methods on three benchmark data sets. The empirical evidence indicates that our model is superior to the considered baselines. Although our method does not back-propagate through the black-box network, it generally surpasses state-of-the-art methods that regard the teacher as a glass-box model. Our code is available at: <https://github.com/antoniobarbalau/black-box-ripper>.

[Bayesian Optimization of Risk Measures](#)

- Sait Cakmak, Raul Astudillo Marban, Peter Frazier, Enlu Zhou
- abstract@[open-review](#): We consider Bayesian optimization of objective functions of the form $\rho[F(x, W)]$, where F is a black-box expensive-to-evaluate function and ρ denotes either the VaR or CVaR risk measure, computed with respect to the randomness induced by the environmental random variable W . Such problems arise in decision making under uncertainty, such as in portfolio optimization and robust systems design. We propose a family of novel Bayesian optimization algorithms that exploit the structure of the objective function to substantially improve sampling efficiency. Instead of modeling the objective function directly as is typical in Bayesian optimization, these algorithms model F as a Gaussian process, and use the implied posterior on the objective function to decide which points to evaluate. We demonstrate the effectiveness of our approach in a variety of numerical experiments.

[TorsionNet: A Reinforcement Learning Approach to Sequential Conformer Search](#)

- Tarun Gogineni, Ziping Xu, Exequiel Punzalan, Runxuan Jiang, Joshua Kammeraad, Ambuj Tewari, Paul Zimmerman
- abstract@[open-review](#): Molecular geometry prediction of flexible molecules, or conformer search, is a long-standing challenge in computational chemistry. This task is of great importance for predicting structure-activity relationships for a wide variety of substances ranging from biomolecules to ubiquitous materials. Substantial computational resources are invested in Monte Carlo and Molecular Dynamics methods to generate diverse and representative conformer sets for medium to large molecules, which are yet intractable to chemoinformatic conformer search methods. We present TorsionNet, an efficient sequential conformer search technique based on reinforcement learning under the rigid rotor approximation. The model is trained via curriculum learning, whose theoretical benefit is explored in detail, to maximize a novel metric grounded in thermodynamics called the Gibbs Score. Our experimental results show that TorsionNet outperforms the highest-scoring chemoinformatics method by 4x on large branched alkanes, and by several orders of magnitude on the previously unexplored biopolymer lignin, with applications in renewable energy. TorsionNet also outperforms the far more exhaustive but computationally intensive Self-Guided Molecular Dynamics sampling method.

[GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis](#)

- Katja Schwarz, Yiyi Liao, Michael Niemeyer, Andreas Geiger
- abstract@[open-review](#): While 2D generative adversarial networks have enabled high-resolution image synthesis, they largely lack an understanding of the 3D world and the image formation process. Thus, they do not provide precise control over camera viewpoint or object pose. To address this problem, several recent approaches leverage intermediate voxel-based representations in combination with differentiable rendering. However, existing methods either produce low image resolution or fall short in disentangling camera and scene properties, e.g., the object identity may vary with the viewpoint. In this paper, we propose a generative model for radiance fields which have recently proven successful for novel view synthesis of a single scene. In contrast to voxel-based representations, radiance fields are not confined to a coarse discretization of the 3D space, yet allow for disentangling camera and scene properties while degrading gracefully in the presence of reconstruction ambiguity. By introducing a multi-scale patch-based discriminator, we demonstrate synthesis of high-resolution images while training our model from unposed 2D images alone. We systematically analyze our approach on several challenging synthetic and real-world datasets. Our experiments reveal that radiance fields are a powerful representation for generative image synthesis, leading to 3D consistent models that render with high fidelity.

[PIE-NET: Parametric Inference of Point Cloud Edges](#)

- Xiaogang Wang, Yuelang Xu, Kai Xu, Andrea Tagliasacchi, Bin Zhou, Ali Mahdavi-Amiri, Hao Zhang
- abstract@[open-review](#): We introduce an end-to-end learnable technique to robustly identify feature edges in 3D point cloud data. We represent these edges as a collection of parametric curves (i.e., lines, circles, and B-splines). Accordingly, our deep neural network, coined PIE-NET, is trained for parametric

inference of edges. The network relies on a "region proposal" architecture, where a first module proposes an over-complete collection of edge and corner points, and a second module ranks each proposal to decide whether it should be considered. We train and evaluate our method on the ABC dataset, a large dataset of CAD models, and compare our results to those produced by traditional (non-learning) processing pipelines, as well as a recent deep learning based edge detector (EC-NET). Our results significantly improve over the state-of-the-art from both a quantitative and qualitative standpoint.

[A Simple Language Model for Task-Oriented Dialogue](#)

- Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, Richard Socher
- abstract@[open-review](#): Task-oriented dialogue is often decomposed into three tasks: understanding user input, deciding actions, and generating a response. While such decomposition might suggest a dedicated model for each sub-task, we find a simple, unified approach leads to state-of-the-art performance on the MultiWOZ dataset. SimpleTOD is a simple approach to task-oriented dialogue that uses a single, causal language model trained on all sub-tasks recast as a single sequence prediction problem. This allows SimpleTOD to fully leverage transfer learning from pre-trained, open domain, causal language models such as GPT-2. SimpleTOD improves over the prior state-of-the-art in joint goal accuracy for dialogue state tracking, and our analysis reveals robustness to noisy annotations in this setting. SimpleTOD also improves the main metrics used to evaluate action decisions and response generation in an end-to-end setting: inform rate by 8.1 points, success rate by 9.7 points, and combined score by 7.2 points.

[A Continuous-Time Mirror Descent Approach to Sparse Phase Retrieval](#)

- Fan Wu, Patrick Rebeschini
- abstract@[open-review](#): We analyze continuous-time mirror descent applied to sparse phase retrieval, which is the problem of recovering sparse signals from a set of magnitude-only measurements. We apply mirror descent to the unconstrained empirical risk minimization problem (batch setting), using the square loss and square measurements. We provide a full convergence analysis of the algorithm in this non-convex setting and prove that, with the hypentropy mirror map, mirror descent recovers any $\$k\$$ -sparse vector $\$\\mathbf{x}^*\\in\\mathbb{R}^n\$$ with minimum (in modulus) non-zero entry on the order of $\$\\sqrt{k}\\$$ Gaussian measurements, modulo logarithmic terms. This yields a simple algorithm which, unlike most existing approaches to sparse phase retrieval, adapts to the sparsity level, without including thresholding steps or adding regularization terms. Our results also provide a principled theoretical understanding for Hadamard Wirtinger flow [54], as Euclidean gradient descent applied to the empirical risk problem with Hadamard parametrization can be recovered as a first-order approximation to mirror descent in discrete time.

[Confidence sequences for sampling without replacement](#)

- Ian Waudby-Smith, Aaditya Ramdas
- abstract@[open-review](#): Many practical tasks involve sampling sequentially without replacement (WoR) from a finite population of size N , in an attempt to estimate some parameter θ^* . Accurately quantifying uncertainty throughout this process is a nontrivial task, but is necessary because it often determines when we stop collecting samples and confidently report a result. We present a suite of tools for designing confidence sequences (CS) for θ^* . A CS is a sequence of confidence sets $(C_n)_{n=1}^N$, that shrink in size, and all contain θ^* simultaneously with high probability. We first exploit a relationship between Bayesian posteriors and martingales to construct a (frequentist) CS for the parameters of a hypergeometric distribution. We then present Hoeffding- and empirical-Bernstein-type time-uniform CSs and fixed-time confidence intervals for sampling WoR which improve on previous bounds in the literature.

[A mean-field analysis of two-player zero-sum games](#)

- Carles Domingo-Enrich, Samy Jelassi, Arthur Mensch, Grant Rotskoff, Joan Bruna
- abstract@[open-review](#): Finding Nash equilibria in two-player zero-sum continuous games is a central problem in machine learning, e.g. for training both GANs and robust models. The existence of pure Nash equilibria requires strong conditions which are not typically met in practice. Mixed Nash equilibria exist in greater generality and may be found using mirror descent. Yet this approach does not scale to high dimensions. To address this limitation, we parametrize mixed strategies as mixtures of particles, whose positions and weights are updated using gradient descent-ascent. We study this dynamics as an interacting gradient flow over measure spaces endowed with the Wasserstein-Fisher-Rao metric. We establish global convergence to an approximate equilibrium for the related Langevin gradient-ascent dynamic. We prove a law of large numbers that relates particle dynamics to mean-field dynamics. Our method identifies mixed equilibria in high dimensions and is demonstrably effective for training mixtures of GANs.

[Leap-Of-Thought: Teaching Pre-Trained Models to Systematically Reason Over Implicit Knowledge](#)

- Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, Jonathan Berant
- abstract@[open-review](#): To what extent can a neural network systematically reason over symbolic facts? Evidence suggests that large pre-trained language models (LMs) acquire some reasoning capacity, but this ability is difficult to control. Recently, it has been shown that Transformer-based models succeed in consistent reasoning over explicit symbolic facts, under a "closed-world" assumption. However, in an open-domain setup, it is desirable to tap into the vast reservoir of implicit knowledge already encoded in the parameters of pre-trained LMs. In this work, we provide a first demonstration that LMs can be trained to reliably perform systematic reasoning combining both implicit, pre-trained knowledge and explicit natural language statements. To do this, we describe a procedure for automatically generating datasets that teach a model new reasoning skills, and demonstrate that models learn to effectively perform inference which involves implicit taxonomic and world knowledge, chaining and counting. Finally, we show that "teaching" models to reason generalizes beyond the training distribution: they successfully compose the usage of multiple reasoning skills in single examples. Our work paves a path towards open-domain systems that constantly improve by interacting with users who can instantly correct a model by adding simple natural language statements.

[Pipeline PSRO: A Scalable Approach for Finding Approximate Nash Equilibria in Large Games](#)

- Stephen McAleer, JB Lanier, Roy Fox, Pierre Baldi
- abstract@[open-review](#): Finding approximate Nash equilibria in zero-sum imperfect-information games is challenging when the number of information states is large. Policy Space Response Oracles (PSRO) is a deep reinforcement learning algorithm grounded in game theory that is guaranteed to converge to an approximate Nash equilibrium. However, PSRO requires training a reinforcement learning policy at each iteration, making it too slow for large games. We show through counterexamples and experiments that DCH and Rectified PSRO, two existing approaches to scaling up PSRO, fail to converge even in small games. We introduce Pipeline PSRO (P2SRO), the first scalable PSRO-based method for finding approximate Nash equilibria in large zero-sum imperfect-information games. P2SRO is able to parallelize PSRO with convergence guarantees by maintaining a hierarchical pipeline of reinforcement learning workers, each training against the policies generated by lower levels in the hierarchy. We show that unlike existing methods, P2SRO converges to an approximate Nash equilibrium, and does so faster as the number of parallel workers increases, across a variety of imperfect information games. We also introduce an open-source environment for Barrage Stratego, a variant of Stratego with an approximate game tree complexity of 10^{50} . P2SRO is able to achieve state-of-the-art performance on Barrage Stratego and beats all existing bots. Experiment code is available at <https://github.com/JBLanier/pipeline-psro>.

[Improving Sparse Vector Technique with Renyi Differential Privacy](#)

- Yuqing Zhu, Yu-Xiang Wang
- abstract@[open-review](#): The Sparse Vector Technique (SVT) is one of the most fundamental algorithmic tools in differential privacy (DP). It also plays a central role in the state-of-the-art algorithms for adaptive data analysis and model-agnostic private learning. In this paper, we revisit SVT from the lens of Renyi differential privacy, which results in new privacy bounds, new theoretical insight and new variants of SVT algorithms. A notable example is a Gaussian mechanism version of SVT, which provides better utility over the standard (Laplace-mechanism-based) version thanks to its more concentrated noise and tighter composition. Extensive empirical evaluation demonstrates the merits of Gaussian SVT over the Laplace SVT and other alternatives, which encouragingly suggests that using Gaussian SVT as a drop-in replacement could make SVT-based algorithms practical in downstream tasks.

[Latent Template Induction with Gumbel-CRFs](#)

- Yao Fu, Chuanqi Tan, Bin Bi, Mosh Chen, Yansong Feng, Alexander Rush
- abstract@[open-review](#): Learning to control the structure of sentences is a challenging problem in text generation. Existing work either relies on simple deterministic approaches or RL-based hard structures. We explore the use of structured variational autoencoders to infer latent templates for sentence generation using a soft, continuous relaxation in order to utilize reparameterization for training. Specifically, we propose a Gumbel-CRF, a continuous relaxation of the CRF sampling algorithm using a relaxed Forward-Filtering Backward-Sampling (FFBS) approach. As a reparameterized gradient estimator, the Gumbel-CRF gives more stable gradients than score-function based estimators. As a structured inference network, we show that it learns interpretable templates during training, which allows us to control the decoder during testing. We demonstrate the effectiveness of our methods with experiments on data-to-text generation and unsupervised paraphrase generation.

[Instance Based Approximations to Profile Maximum Likelihood](#)

- Nima Anari, Moses Charikar, Kirankumar Shiragur, Aaron Sidford
- abstract@[open-review](#): In this paper we provide a new efficient algorithm for approximately computing the profile maximum likelihood (PML) distribution, a prominent quantity in symmetric property estimation. We provide an algorithm which matches the previous best known efficient algorithms for computing approximate PML distributions and improves when the number of distinct observed frequencies in the given instance is small. We achieve this result by exploiting new sparsity structure in approximate PML distributions and providing a new matrix rounding algorithm, of independent interest. Leveraging this result, we obtain the first provable computationally efficient implementation of PseudoPML, a general framework for estimating a broad class of symmetric properties. Additionally, we obtain efficient PML-based estimators for distributions with small profile entropy, a natural instance-based complexity measure. Further, we provide a simpler and more practical PseudoPML implementation that matches the best-known theoretical guarantees of such an estimator and evaluate this method empirically.

[Factorizable Graph Convolutional Networks](#)

- Yiding Yang, Zunlei Feng, Mingli Song, Xinchao Wang
- abstract@[open-review](#): Graphs have been widely adopted to denote structural connections between entities. The relations are in many cases heterogeneous, but entangled together and denoted merely as a single edge between a pair of nodes. For example, in a social network graph, users in different latent relationships like friends and colleagues, are usually connected via a bare edge that conceals such intrinsic connections. In this paper, we introduce a novel graph convolutional network (GCN), termed as factorizable graph convolutional network (FactorGCN), that explicitly disentangles such intertwined relations encoded in a graph. FactorGCN takes a simple graph as input, and disentangles it into several factorized graphs, each of which represents a latent and disentangled relation among nodes. The features of the nodes are then aggregated separately in each factorized latent space to produce disentangled features, which further leads to better performances for downstream tasks. We evaluate the proposed FactorGCN both qualitatively and quantitatively on the synthetic and real-world datasets, and demonstrate that it yields truly encouraging results in terms of both disentangling and feature aggregation. Code is publicly available at <https://github.com/ihollywhy/FactorGCN.PyTorch>.

[Guided Adversarial Attack for Evaluating and Enhancing Adversarial Defenses](#)

- Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, Venkatesh Babu R
- abstract@[open-review](#): Advances in the development of adversarial attacks have been fundamental to the progress of adversarial defense research. Efficient and effective attacks are crucial for reliable evaluation of defenses, and also for developing robust models. Adversarial attacks are often generated by maximizing standard losses such as the cross-entropy loss or maximum-margin loss within a constraint set using Projected Gradient Descent (PGD). In this work, we introduce a relaxation term to the standard loss, that finds more suitable gradient-directions, increases attack efficacy and leads to more efficient adversarial training. We propose Guided Adversarial Margin Attack (GAMA), which utilizes function mapping of the clean image to guide the generation of adversaries, thereby resulting in stronger attacks. We evaluate our attack against multiple defenses and show improved performance when compared to existing attacks. Further, we propose Guided Adversarial Training (GAT), which achieves state-of-the-art performance amongst single-step defenses by utilizing the proposed relaxation term for both attack generation and training.

[A Study on Encodings for Neural Architecture Search](#)

- Colin White, Willie Neiswanger, Sam Nolen, Yash Savani
- abstract@[open-review](#): Neural architecture search (NAS) has been extensively studied in the past few years. A popular approach is to represent each neural architecture in the search space as a directed acyclic graph (DAG), and then search over all DAGs by encoding the adjacency matrix and list of operations as a set of hyperparameters. Recent work has demonstrated that even small changes to the way each architecture is encoded can have a significant effect on the performance of NAS algorithms. In this work, we present the first formal study on the effect of architecture encodings for NAS, including a theoretical grounding and an empirical study. First we formally define architecture encodings and give a theoretical characterization on the scalability of the encodings we study. Then we identify the main encoding-dependent subroutines which NAS algorithms employ, running experiments to show which encodings work best with each subroutine for many popular algorithms. The experiments act as an ablation study for prior work, disentangling the algorithmic and encoding-based contributions, as well as a guideline for future work. Our results demonstrate that NAS encodings are an important design decision which can have a significant impact on overall performance. Our code is available at <https://github.com/naszilla/naszilla>.

[Noise2Same: Optimizing A Self-Supervised Bound for Image Denoising](#)

- Yaochen Xie, Zhengyang Wang, Shuiwang Ji
- abstract@[open-review](#): Self-supervised frameworks that learn denoising models with merely individual noisy images have shown strong capability and promising performance in various image denoising tasks. Existing self-supervised denoising frameworks are mostly built upon the same theoretical foundation, where the denoising models are required to be J-invariant. However, our analyses indicate that the current theory and the J-invariance may lead to denoising models with reduced performance. In this work, we introduce Noise2Same, a novel self-supervised denoising framework. In Noise2Same, a new self-supervised loss is proposed by deriving a self-supervised upper bound of the typical supervised loss. In particular, Noise2Same requires neither J-invariance nor extra information about the noise model and can be used in a wider range of denoising applications. We analyze our proposed Noise2Same both theoretically and experimentally. The experimental results show that our Noise2Same remarkably outperforms previous self-supervised denoising methods in terms of denoising performance and training efficiency.

[Early-Learning Regularization Prevents Memorization of Noisy Labels](#)

- Sheng Liu, Jonathan Niles-Weed, Narges Razavian, Carlos Fernandez-Granda
- abstract@[open-review](#): We propose a novel framework to perform classification via deep learning in the presence of noisy annotations. When trained on noisy labels, deep neural networks have been observed to first fit the training data with clean labels during an "early learning" phase, before eventually memorizing the examples with false labels. We prove that early learning and memorization are fundamental phenomena in high-dimensional classification tasks, even in simple linear models, and give a theoretical explanation in this setting. Motivated by these findings, we develop a new technique for noisy classification tasks, which exploits the progress of the early learning phase. In contrast with existing approaches, which use the model output during early learning to detect the examples with clean labels, and either ignore or attempt to correct the false labels, we take a different route and instead capitalize on early learning via regularization. There are two key elements to our approach. First, we leverage semi-supervised learning techniques to produce target probabilities based on the model outputs. Second, we design a regularization term that steers the model towards these targets, implicitly preventing memorization of the false labels. The resulting framework is shown to provide robustness to noisy annotations on several standard benchmarks and real-world datasets, where it achieves results comparable to the state of the art.

[LAPAR: Linearly-Assembled Pixel-Adaptive Regression Network for Single Image Super-resolution and Beyond](#)

- Wenbo Li, Kun Zhou, Lu Qi, Nianjuan Jiang, Jiangbo Lu, Jiaya Jia
- abstract@[open-review](#): Single image super-resolution (SISR) deals with a fundamental problem of upsampling a low-resolution (LR) image to its high-resolution (HR) version. Last few years have witnessed impressive progress propelled by deep learning methods. However, one critical challenge faced by existing methods is to strike a sweet spot of deep model complexity and resulting SISR quality. This paper addresses this pain point by proposing a linearly-assembled pixel-adaptive regression network (LAPAR), which casts the direct LR to HR mapping learning into a linear coefficient regression task over a dictionary of multiple predefined filter bases. Such a parametric representation renders our model highly lightweight and easy to optimize while achieving state-of-the-art results on SISR benchmarks. Moreover, based on the same idea, LAPAR is extended to tackle other restoration tasks, e.g., image denoising and JPEG image deblocking, and again, yields strong performance.

[Learning Parities with Neural Networks](#)

- Amit Daniely, Eran Malach
- abstract@[open-review](#): In recent years we see a rapidly growing line of research which shows learnability of various models via common neural network algorithms. Yet, besides a very few outliers, these results show learnability of models that can be learned using linear methods. Namely, such results show that learning neural-networks with gradient-descent is competitive with learning a linear classifier on top of a data-independent representation of the examples. This leaves much to be desired, as neural networks are far more successful than linear methods. Furthermore, on the more conceptual level, linear models don't seem to capture the ``deepness'' of deep networks. In this paper we make a step towards showing learnability of models that are inherently non-linear. We show that under certain distributions, sparse parities are learnable via gradient descent on depth-two network. On the other hand, under the same distributions, these parities cannot be learned efficiently by linear methods.

[Consistent Plug-in Classifiers for Complex Objectives and Constraints](#)

- Shiv Kumar Tavker, Harish Guruprasad Ramaswamy, Harikrishna Narasimhan
- abstract@[open-review](#): We present a statistically consistent algorithm for constrained classification problems where the objective (e.g. F-measure, G-mean) and the constraints (e.g. demographic parity, coverage) are defined by general functions of the confusion matrix. The key idea is to reduce the problem into a sequence of plug-in classifier learning problems, which is done by formulating an optimization problem over the intersection of the set of achievable confusion matrices and the set of feasible matrices. For objective and constraints that are convex functions of the confusion matrix, our algorithm requires $\mathcal{O}(1/\epsilon^2)$ calls to the plug-in routine, which improves on the $\mathcal{O}(1/\epsilon^3)$ rate achieved by Narasimhan (2018). We demonstrate empirically that our algorithm performs at least as well as the state-of-the-art methods for these problems.

[Movement Pruning: Adaptive Sparsity by Fine-Tuning](#)

- Victor Sanh, Thomas Wolf, Alexander Rush
- abstract@[open-review](#): Magnitude pruning is a widely used strategy for reducing model size in pure supervised learning; however, it is less effective in the transfer learning regime that has become standard for state-of-the-art natural language processing applications. We propose the use of movement pruning, a simple, deterministic first-order weight pruning method that is more adaptive to pretrained model fine-tuning. We give mathematical foundations to the method and compare it to existing zeroth- and first-order pruning methods. Experiments show that when pruning large pretrained language models, movement pruning shows significant improvements in high-sparsity regimes. When combined with distillation, the approach achieves minimal accuracy loss with down to only 3% of the model parameters.

[Sanity-Checking Pruning Methods: Random Tickets can Win the Jackpot](#)

- Jingtong Su, Yihang Chen, Tianle Cai, Tianhao Wu, Ruiqi Gao, Liwei Wang, Jason D. Lee
- abstract@[open-review](#): Network pruning is a method for reducing test-time computational resource requirements with minimal performance degradation. Conventional wisdom of pruning algorithms suggests that: (1) Pruning methods exploit information from training data to find good subnetworks; (2) The architecture of the pruned network is crucial for good performance. In this paper, we conduct sanity checks for the above beliefs on several recent unstructured pruning methods and surprisingly find that: (1) A set of methods which aims to find good subnetworks of the randomly-initialized network (which we call initial tickets"), hardly exploits any information from the training data; (2) For the pruned networks obtained by these methods, randomly changing the preserved weights in each layer, while keeping the total number of preserved weights unchanged per layer, does not affect the final performance. These findings inspire us to choose a series of simple \text{data-independent} prune ratios for each layer, and randomly prune each layer accordingly to get a subnetwork (which we call random tickets"). Experimental results show that our zero-shot random tickets outperforms or attains similar performance compared to existing initial tickets". In addition, we identify one existing pruning method that passes our sanity checks. We hybridize the ratios in our random ticket with this method and propose a new method called hybrid tickets", which achieves further improvement.

[Online Matrix Completion with Side Information](#)

- Mark Herbster, Stephen Pasteris, Lisa Tse
- abstract@[open-review](#): We give an online algorithm and prove novel mistake and regret bounds for online binary matrix completion with side information. The mistake bounds we prove are of the form $\tilde{O}(\tilde{D}/\gamma^2)$. The term $1/\gamma^2$ is analogous to the usual margin term in SVM (perceptron) bounds. More specifically, if we assume that there is some factorization of the underlying $m \times n$ matrix into PQ^T , where the rows of P are interpreted as "classifiers" in R^d and the rows of Q as "instances" in R^d , then γ is the maximum (normalized) margin over all factorizations PQ^T consistent with the observed matrix. The quasi-dimension term D measures the quality of side information. In the presence of vacuous side information, $D = m+n$. However, if the side information is predictive of the underlying factorization of the matrix, then in an ideal case, $D \in O(k+l)$ where k is the number of distinct row factors and l is the number of distinct column factors. We additionally provide a generalization of our algorithm to the inductive

setting. In this setting, we provide an example where the side information is not directly specified in advance. For this example, the quasi-dimension D is now bounded by $O(k^2 + l^2)$.

[Position-based Scaled Gradient for Model Quantization and Pruning](#)

- Jangho Kim, KiYoon Yoo, Nojun Kwak
- abstract@[open-review](#): We propose the position-based scaled gradient (PSG) that scales the gradient depending on the position of a weight vector to make it more compression-friendly. First, we theoretically show that applying PSG to the standard gradient descent (GD), which is called PSGD, is equivalent to the GD in the warped weight space, a space made by warping the original weight space via an appropriately designed invertible function. Second, we empirically show that PSG acting as a regularizer to a weight vector is favorable for model compression domains such as quantization and pruning. PSG reduces the gap between the weight distributions of a full-precision model and its compressed counterpart. This enables the versatile deployment of a model either as an uncompressed mode or as a compressed mode depending on the availability of resources. The experimental results on CIFAR-10/100 and ImageNet datasets show the effectiveness of the proposed PSG in both domains of pruning and quantization even for extremely low bits. The code is released in Github.

[Online Learning with Primary and Secondary Losses](#)

- Avrim Blum, Han Shao
- abstract@[open-review](#): We study the problem of online learning with primary and secondary losses. For example, a recruiter making decisions of which job applicants to hire might weigh false positives and false negatives equally (the primary loss) but the applicants might weigh false negatives much higher (the secondary loss). We consider the following question: Can we combine ``expert advice'' to achieve low regret with respect to the primary loss, while at the same time performing $\{\text{em not much worse than the worst expert}\}$ with respect to the secondary loss? Unfortunately, we show that this goal is unachievable without any bounded variance assumption on the secondary loss. More generally, we consider the goal of minimizing the regret with respect to the primary loss and bounding the secondary loss by a linear threshold. On the positive side, we show that running any switching-limited algorithm can achieve this goal if all experts satisfy the assumption that the secondary loss does not exceed the linear threshold by $\$o(T)\$$ for any time interval. If not all experts satisfy this assumption, our algorithms can achieve this goal given access to some external oracles which determine when to deactivate and reactivate experts.

[Graph Information Bottleneck](#)

- Tailin Wu, Hongyu Ren, Pan Li, Jure Leskovec
- abstract@[open-review](#): Representation learning of graph-structured data is challenging because both graph structure and node features carry important information. Graph Neural Networks (GNNs) provide an expressive way to fuse information from network structure and node features. However, GNNs are prone to adversarial attacks. Here we introduce Graph Information Bottleneck (GIB), an information-theoretic principle that optimally balances expressiveness and robustness of the learned representation of graph-structured data. Inheriting from the general Information Bottleneck (IB), GIB aims to learn the minimal sufficient representation for a given task by maximizing the mutual information between the representation and the target, and simultaneously constraining the mutual information between the representation and the input data. Different from the general IB, GIB regularizes the structural as well as the feature information. We design two sampling algorithms for structural regularization and instantiate the GIB principle with two new models: GIB-Cat and GIB-Bern, and demonstrate the benefits by evaluating the resilience to adversarial attacks. We show that our proposed models are more robust than state-of-the-art graph defense models. GIB-based models empirically achieve up to 31% improvement with adversarial perturbation of the graph structure as well as node features.

[The Complexity of Adversarially Robust Proper Learning of Halfspaces with Agnostic Noise](#)

- Ilias Diakonikolas, Daniel M. Kane, Pasin Manurangsi
- abstract@[open-review](#): We study the computational complexity of adversarially robust proper learning of halfspaces in the distribution-independent agnostic PAC model, with a focus on L_p perturbations. We give a computationally efficient learning algorithm and a nearly matching computational hardness result for this problem. An interesting implication of our findings is that the L_∞ perturbations case is provably computationally harder than the case $2 \leq p < \infty$.

[Adaptive Online Estimation of Piecewise Polynomial Trends](#)

- Dheeraj Baby, Yu-Xiang Wang
- abstract@[open-review](#): We consider the framework of non-stationary stochastic optimization [Besbes et.al. 2015] with squared error losses and noisy gradient feedback where the dynamic regret of an online learner against a time varying comparator sequence is studied. Motivated from the theory of non-parametric regression, we introduce a *new variational constraint* that enforces the comparator sequence to belong to a discrete k^{th} order Total Variation ball of radius C_n . This variational constraint models comparators that have piece-wise polynomial structure which has many relevant practical applications [Tibshirani2015]. By establishing connections to the theory of wavelet based non-parametric regression, we design a *polynomial time* algorithm that achieves the nearly *optimal dynamic regret* of $\tilde{O}(n^{\frac{1}{2k+3}}C_n^{\frac{2}{2k+3}})$. The proposed policy is *adaptive to the unknown radius* C_n . Further, we show that the same policy is minimax optimal for several other non-parametric families of interest.

[RNNPool: Efficient Non-linear Pooling for RAM Constrained Inference](#)

- Oindrila Saha, Aditya Kusupati, Harsha Vardhan Simhadri, Manik Varma, Prateek Jain
- abstract@[open-review](#): Standard Convolutional Neural Networks (CNNs) designed for computer vision tasks tend to have large intermediate activation maps. These require large working memory and are thus unsuitable for deployment on resource-constrained devices typically used for inference on the edge. Aggressively downsampling the images via pooling or strided convolutions can address the problem but leads to a significant decrease in accuracy due to gross aggregation of the feature map by standard pooling operators. In this paper, we introduce RNNPool, a novel pooling operator based on Recurrent Neural Networks (RNNs), that efficiently aggregates features over large patches of an image and rapidly downsamples activation maps. Empirical evaluation indicates that an RNNPool layer can effectively replace multiple blocks in a variety of architectures such as MobileNets, DenseNet when applied to standard vision tasks like image classification and face detection. That is, RNNPool can significantly decrease computational complexity and peak memory usage for inference while retaining comparable accuracy. We use RNNPool with the standard S3FD architecture to construct a face detection method that achieves state-of-the-art MAP for tiny ARM Cortex-M4 class microcontrollers with under 256 KB of RAM. Code is released at <https://github.com/Microsoft/EdgeML>.

[Agnostic Learning with Multiple Objectives](#)

- Corinna Cortes, Mehryar Mohri, Javier Gonzalvo, Dmitry Storcheus
- abstract@[open-review](#): Most machine learning tasks are inherently multi-objective. This means that the learner has to come up with a model that performs well across a number of base objectives $\{L_1, \dots, L_p\}$, as opposed to a single one. Since optimizing with respect to multiple objectives at the

same time is often computationally expensive, the base objectives are often combined in an ensemble $\sum_{k=1}^p \lambda_k c_L(k)$, thereby reducing the problem to scalar optimization. The mixture weights λ_k are set to uniform or some other fixed distribution, based on the learner's preferences. We argue that learning with a fixed distribution on the mixture weights runs the risk of overfitting to some individual objectives and significantly harming others, despite performing well on an entire ensemble. Moreover, in reality, the true preferences of a learner across multiple objectives are often unknown or hard to express as a specific distribution. Instead, we propose a new framework of Agnostic Learning with Multiple Objectives (ALMO), where a model is optimized for λ weights in the mixture of base objectives. We present data-dependent Rademacher complexity guarantees for learning in the ALMO framework, which are used to guide a scalable optimization algorithm and the corresponding regularization. We present convergence guarantees for this algorithm, assuming convexity of the loss functions and the underlying hypothesis space. We further implement the algorithm in a popular symbolic gradient computation framework and empirically demonstrate on a number of datasets the benefits of ALMO framework versus learning with a fixed mixture weights distribution.

[3D Multi-bodies: Fitting Sets of Plausible 3D Human Models to Ambiguous Image Data](#)

- Benjamin Biggs, David Novotny, Sébastien Ehrhardt, Hanbyul Joo, Ben Graham, Andrea Vedaldi
- abstract@[open-review](#): We consider the problem of obtaining dense 3D reconstructions of deformable objects from single and partially occluded views. In such cases, the visual evidence is usually insufficient to identify a 3D reconstruction uniquely, so we aim at recovering several plausible reconstructions compatible with the input data. We suggest that ambiguities can be modeled more effectively by parametrizing the possible body shapes and poses via a suitable 3D model, such as SMPL for humans. We propose to learn a multi-hypothesis neural network regressor using a best-of-M loss, where each of the M hypotheses is constrained to lie on a manifold of plausible human poses by means of a generative model. We show that our method outperforms alternative approaches in ambiguous pose recovery on standard benchmarks for 3D humans, and in heavily occluded versions of these benchmarks.

[Auto-Panoptic: Cooperative Multi-Component Architecture Search for Panoptic Segmentation](#)

- Yangxin Wu, Gengwei Zhang, Hang Xu, Xiaodan Liang, Liang Lin
- abstract@[open-review](#): Panoptic segmentation is posed as a new popular test-bed for the state-of-the-art holistic scene understanding methods with the requirement of simultaneously segmenting both foreground things and background stuff. The state-of-the-art panoptic segmentation network exhibits high structural complexity in different network components, i.e. backbone, proposal-based foreground branch, segmentation-based background branch, and feature fusion module across branches, which heavily relies on expert knowledge and tedious trials. In this work, we propose an efficient, cooperative and highly automated framework to simultaneously search for all main components including backbone, segmentation branches, and feature fusion module in a unified panoptic segmentation pipeline based on the prevailing one-shot Network Architecture Search (NAS) paradigm. Notably, we extend the common single-task NAS into the multi-component scenario by taking the advantages of the newly proposed intra-modular search space and problem-oriented inter-modular search space, which helps us to obtain an optimal network architecture that not only performs well in both instance segmentation and semantic segmentation tasks but also be aware of the reciprocal relations between foreground things and background stuff classes. To relieve the vast computation burden incurred by applying NAS to complicated network architectures, we present a novel path-priority greedy search policy to find a robust, transferrable architecture with significantly reduced searching overhead. Our searched architecture, namely Auto-Panoptic, achieves the new state-of-the-art on the challenging COCO and ADE20K benchmarks. Moreover, extensive experiments are conducted to demonstrate the effectiveness of path-priority policy and transferability of Auto-Panoptic across different datasets.

[Differentiable Top-k with Optimal Transport](#)

- Yujia Xie, Hanjun Dai, Minshuo Chen, Bo Dai, Tuo Zhao, Hongyuan Zha, Wei Wei, Tomas Pfister
- abstract@[open-review](#): Finding the k largest or smallest elements from a collection of scores, i.e., top-k operation, is an important model component widely used in information retrieval, machine learning, and data mining. However, if the top-k operation is implemented in an algorithmic way, e.g., using bubble algorithm, the resulted model cannot be trained in an end-to-end way using prevalent gradient descent algorithms. This is because these implementations typically involve swapping indices, whose gradient cannot be computed. Moreover, the corresponding mapping from the input scores to the indicator vector of whether this element belongs to the top-k set is essentially discontinuous. To address the issue, we propose a smoothed approximation, namely SOFT (Scalable Optimal transport-based diFferenTiable) top-k operator. Specifically, our SOFT top-k operator approximates the output of top-k operation as the solution of an Entropic Optimal Transport (EOT) problem. The gradient of the SOFT operator can then be efficiently approximated based on the optimality conditions of EOT problem. We then apply the proposed operator to k-nearest neighbors algorithm and beam search algorithm. The numerical experiment demonstrates their achieve improved performance.

[Information-theoretic Task Selection for Meta-Reinforcement Learning](#)

- Ricardo Luna Gutierrez, Matteo Leonetti
- abstract@[open-review](#): In Meta-Reinforcement Learning (meta-RL) an agent is trained on a set of tasks to prepare for and learn faster in new, unseen, but related tasks. The training tasks are usually hand-crafted to be representative of the expected distribution of target tasks and hence all used in training. We show that given a set of training tasks, learning can be both faster and more effective (leading to better performance in the target tasks), if the training tasks are appropriately selected. We propose a task selection algorithm based on information theory, which optimizes the set of tasks used for training in meta-RL, irrespectively of how they are generated. The algorithm establishes which training tasks are both sufficiently relevant for the target tasks, and different enough from one another. We reproduce different meta-RL experiments from the literature and show that our task selection algorithm improves the final performance in all of them.

[A Limitation of the PAC-Bayes Framework](#)

- Roi Livni, Shay Moran
- abstract@[open-review](#): PAC-Bayes is a useful framework for deriving generalization bounds which was introduced by McAllester ('98). This framework has the flexibility of deriving distribution- and algorithm-dependent bounds, which are often tighter than VC-related uniform convergence bounds.

In this manuscript we present a limitation for the PAC-Bayes framework. We demonstrate an easy learning task which is not amenable to a PAC-Bayes analysis.

Specifically, we consider the task of linear classification in 1D; it is well-known that this task is learnable using just $O(\log(1/\delta)/\epsilon)$ examples. On the other hand, we show that this fact can not be proved using a PAC-Bayes analysis: for any algorithm that learns 1-dimensional linear classifiers there exists a (realizable) distribution for which the PAC-Bayes bound is arbitrarily large.

[On Completeness-aware Concept-Based Explanations in Deep Neural Networks](#)

- Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, Pradeep Ravikumar
- abstract@[open-review](#): Human explanations of high-level decisions are often expressed in terms of key concepts the decisions are based on. In this paper, we study such concept-based explainability for Deep Neural Networks (DNNs). First, we define the notion of completeness, which quantifies how sufficient a particular set of concepts is in explaining a model's prediction behavior based on the assumption that complete concept scores are sufficient statistics of the model prediction. Next, we propose a concept discovery method that aims to infer a complete set of concepts that are additionally encouraged to be interpretable, which addresses the limitations of existing methods on concept explanations. To define an importance score for each

discovered concept, we adapt game-theoretic notions to aggregate over sets and propose \text{ConceptSHAP}. Via proposed metrics and user studies, on a synthetic dataset with apriori-known concept explanations, as well as on real-world image and language datasets, we validate the effectiveness of our method in finding concepts that are both complete in explaining the decisions and interpretable.

[Stochastic Recursive Gradient Descent Ascent for Stochastic Nonconvex-Strongly-Concave Minimax Problems](#)

- Luo Luo, Haishan Ye, Zhichao Huang, Tong Zhang
- abstract@[open-review](#): We consider nonconvex-concave minimax optimization problems of the form $\min_{\{\mathbf{x}\}} \max_{\{\mathbf{y} \in \mathcal{Y}\}} f(\{\mathbf{x}\}, \{\mathbf{y}\})$, where f is strongly-concave in \mathbf{y} but possibly nonconvex in \mathbf{x} and \mathcal{Y} is a convex and compact set. We focus on the stochastic setting, where we can only access an unbiased stochastic gradient estimate of f at each iteration. This formulation includes many machine learning applications as special cases such as robust optimization and adversary training. We are interested in finding an $\mathcal{O}(\varepsilon)$ -stationary point of the function $\Phi(\cdot) = \max_{\{\mathbf{y} \in \mathcal{Y}\}} f(\cdot, \{\mathbf{y}\})$. The most popular algorithm to solve this problem is stochastic gradient decent ascent, which requires $\mathcal{O}(\kappa^3 \varepsilon^{-4})$ stochastic gradient evaluations, where κ is the condition number. In this paper, we propose a novel method called Stochastic Recursive gradiEnt Descent Ascent (SREDA), which estimates gradients more efficiently using variance reduction. This method achieves the best known stochastic gradient complexity of $\mathcal{O}(\kappa^3 \varepsilon^{-3})$, and its dependency on ε is optimal for this problem.

[Why Normalizing Flows Fail to Detect Out-of-Distribution Data](#)

- Polina Kirichenko, Pavel Izmailov, Andrew G. Wilson
- abstract@[open-review](#): Detecting out-of-distribution (OOD) data is crucial for robust machine learning systems. Normalizing flows are flexible deep generative models that often surprisingly fail to distinguish between in- and out-of-distribution data: a flow trained on pictures of clothing assigns higher likelihood to handwritten digits. We investigate why normalizing flows perform poorly for OOD detection. We demonstrate that flows learn local pixel correlations and generic image-to-latent-space transformations which are not specific to the target image datasets, focusing on flows based on coupling layers. We show that by modifying the architecture of flow coupling layers we can bias the flow towards learning the semantic structure of the target data, improving OOD detection. Our investigation reveals that properties that enable flows to generate high-fidelity images can have a detrimental effect on OOD detection.

[Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay](#)

- Joao Marques-Silva, Thomas Gerspacher, Martin Cooper, Alexey Ignatiev, Nina Narodytska
- abstract@[open-review](#): Recent work proposed the computation of so-called PI-explanations of Naive Bayes Classifiers (NBCs). PI-explanations are subset-minimal sets of feature-value pairs that are sufficient for the prediction, and have been computed with state-of-the-art exact algorithms that are worst-case exponential in time and space. In contrast, we show that the computation of one PI-explanation for an NBC can be achieved in log-linear time, and that the same result also applies to the more general class of linear classifiers. Furthermore, we show that the enumeration of PI-explanations can be obtained with polynomial delay. Experimental results demonstrate the performance gains of the new algorithms when compared with earlier work. The experimental results also investigate ways to measure the quality of heuristic explanations.

[Unsupervised Translation of Programming Languages](#)

- Baptiste Roziere, Marie-Anne Lachaux, Lowik Chanussot, Guillaume Lample
- abstract@[open-review](#): A transcompiler, also known as source-to-source translator, is a system that converts source code from a high-level programming language (such as C++ or Python) to another. Transcompilers are primarily used for interoperability, and to port codebases written in an obsolete or deprecated language (e.g. COBOL, Python 2) to a modern one. They typically rely on handcrafted rewrite rules, applied to the source code abstract syntax tree. Unfortunately, the resulting translations often lack readability, fail to respect the target language conventions, and require manual modifications in order to work properly. The overall translation process is time-consuming and requires expertise in both the source and target languages, making code-translation projects expensive. Although neural models significantly outperform their rule-based counterparts in the context of natural language translation, their applications to transcompilation have been limited due to the scarcity of parallel data in this domain. In this paper, we propose to leverage recent approaches in unsupervised machine translation to train a fully unsupervised neural transcompiler. We train our model on source code from open source GitHub projects, and show that it can translate functions between C++, Java, and Python with high accuracy. Our method relies exclusively on monolingual source code, requires no expertise in the source or target languages, and can easily be generalized to other programming languages. We also build and release a test set composed of 852 parallel functions, along with unit tests to check the correctness of translations. We show that our model outperforms rule-based commercial baselines by a significant margin.

[Adversarial Style Mining for One-Shot Unsupervised Domain Adaptation](#)

- Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, Yi Yang
- abstract@[open-review](#): We aim at the problem named One-Shot Unsupervised Domain Adaptation. Unlike traditional Unsupervised Domain Adaptation, it assumes that only one unlabeled target sample can be available when learning to adapt. This setting is realistic but more challenging, in which conventional adaptation approaches are prone to failure due to the scarce of unlabeled target data. To this end, we propose a novel Adversarial Style Mining approach, which combines the style transfer module and task-specific module into an adversarial manner. Specifically, the style transfer module iteratively searches for harder stylized images around the one-shot target sample according to the current learning state, leading the task model to explore the potential styles that are difficult to solve in the almost unseen target domain, thus boosting the adaptation performance in a data-scarce scenario. The adversarial learning framework makes the style transfer module and task-specific module benefit each other during the competition. Extensive experiments on both cross-domain classification and segmentation benchmarks verify that ASM achieves state-of-the-art adaptation performance under the challenging one-shot setting.

[Optimally Deceiving a Learning Leader in Stackelberg Games](#)

- Georgios Birmpas, Jiarui Gan, Alexandros Hollender, Francisco Marmolejo, Ninad Rajgopal, Alexandros Voudouris
- abstract@[open-review](#): Recent results in the ML community have revealed that learning algorithms used to compute the optimal strategy for the leader to commit to in a Stackelberg game, are susceptible to manipulation by the follower. Such a learning algorithm operates by querying the best responses or the payoffs of the follower, who consequently can deceive the algorithm by responding as if their payoffs were much different than what they actually are. For this strategic behavior to be successful, the main challenge faced by the follower is to pinpoint the payoffs that would make the learning algorithm compute a commitment so that best responding to it maximizes the follower's utility, according to the true payoffs. While this problem has been considered before, the related literature only focused on the simplified scenario in which the payoff space is finite, thus leaving the general version of the problem unanswered. In this paper, we fill this gap by showing that it is always possible for the follower to efficiently compute (near-)optimal payoffs for various scenarios of learning interaction between the leader and the follower.

[Online Optimization with Memory and Competitive Control](#)

- Guanya Shi, Yiheng Lin, Soon-Jo Chung, Yisong Yue, Adam Wierman
- abstract@[open-review](#): This paper presents competitive algorithms for a novel class of online optimization problems with memory. We consider a setting where the learner seeks to minimize the sum of a hitting cost and a switching cost that depends on the previous \$p\$ decisions. This setting generalizes Smoothed Online Convex Optimization. The proposed approach, Optimistic Regularized Online Balanced Descent, achieves a constant, dimension-free competitive ratio. Further, we show a connection between online optimization with memory and online control with adversarial disturbances. This connection, in turn, leads to a new constant-competitive policy for a rich class of online control problems.

[IDEAL: Inexact DEcentralized Accelerated Augmented Lagrangian Method](#)

- Yossi Arjevani, Joan Bruna, Bugra Can, Mert Gurbuzbalaban, Stefanie Jegelka, Hongzhou Lin
- abstract@[open-review](#): We introduce a framework for designing primal methods under the decentralized optimization setting where local functions are smooth and strongly convex. Our approach consists of approximately solving a sequence of sub-problems induced by the accelerated augmented Lagrangian method, thereby providing a systematic way for deriving several well-known decentralized algorithms including EXTRA and SSDA. When coupled with accelerated gradient descent, our framework yields a novel primal algorithm whose convergence rate is optimal and matched by recently derived lower bounds. We provide experimental results that demonstrate the effectiveness of the proposed algorithm on highly ill-conditioned problems.

[Evolving Graphical Planner: Contextual Global Planning for Vision-and-Language Navigation](#)

- Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky
- abstract@[open-review](#): The ability to perform effective planning is crucial for building an instruction-following agent. When navigating through a new environment, an agent is challenged with (1) connecting the natural language instructions with its progressively growing knowledge of the world; and (2) performing long-range planning and decision making in the form of effective exploration and error correction. Current methods are still limited on both fronts despite extensive efforts. In this paper, we introduce Evolving Graphical Planner (EGP), a module that allows global planning for navigation based on raw sensory input. The module dynamically constructs a graphical representation, generalizes the local action space to allow for more flexible decision making, and performs efficient planning on a proxy representation. We demonstrate our model on a challenging Vision-and-Language Navigation (VLN) task with photorealistic images, and achieve superior performance compared to previous navigation architectures. Concretely, we achieve 53% success rate on the test split of Room-to-Room navigation task (Anderson et al.) through pure imitation learning, outperforming previous architectures by up to 5%.

[Learning from Failure: De-biasing Classifier from Biased Classifier](#)

- Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, Jinwoo Shin
- abstract@[open-review](#): Neural networks often learn to make predictions that overly rely on spurious correlation existing in the dataset, which causes the model to be biased. While previous work tackles this issue by using explicit labeling on the spuriously correlated attributes or presuming a particular bias type, we instead utilize a cheaper, yet generic form of human knowledge, which can be widely applicable to various types of bias. We first observe that neural networks learn to rely on the spurious correlation only when it is easier to learn than the desired knowledge, and such reliance is most prominent during the early phase of training. Based on the observations, we propose a failure-based debiasing scheme by training a pair of neural networks simultaneously. Our main idea is twofold; (a) we intentionally train the first network to be biased by repeatedly amplifying its prejudice, and (b) we debias the training of the second network by focusing on samples that go against the prejudice of the biased network in (a). Extensive experiments demonstrate that our method significantly improves the training of network against various types of biases in both synthetic and real-world datasets. Surprisingly, our framework even occasionally outperforms the debiasing methods requiring explicit supervision of the spuriously correlated attributes.

[Likelihood Regret: An Out-of-Distribution Detection Score For Variational Auto-encoder](#)

- Zhisheng Xiao, Qing Yan, Yali Amit
- abstract@[open-review](#): Deep probabilistic generative models enable modeling the likelihoods of very high dimensional data. An important application of generative modeling should be the ability to detect out-of-distribution (OOD) samples by setting a threshold on the likelihood. However, a recent study shows that probabilistic generative models can, in some cases, assign higher likelihoods on certain types of OOD samples, making the OOD detection rules based on likelihood threshold problematic. To address this issue, several OOD detection methods have been proposed for deep generative models. In this paper, we make the observation that some of these methods fail when applied to generative models based on Variational Auto-encoders (VAE). As an alternative, we propose Likelihood Regret, an efficient OOD score for VAEs. We benchmark our proposed method over existing approaches, and empirical results suggest that our method obtains the best overall OOD detection performances compared with other OOD method applied on VAE.

[Deep Diffusion-Invariant Wasserstein Distributional Classification](#)

- Sung Woo Park, Dong Wook Shu, Junseok Kwon
- abstract@[open-review](#): In this paper, we present a novel classification method called deep diffusion-invariant Wasserstein distributional classification (DeepWDC). DeepWDC represents input data and labels as probability measures to address severe perturbations in input data. It can output the optimal label measure in terms of diffusion invariance, where the label measure is stationary over time and becomes equivalent to a Gaussian measure. Furthermore, DeepWDC minimizes the 2-Wasserstein distance between the optimal label measure and Gaussian measure, which reduces the Wasserstein uncertainty. Experimental results demonstrate that DeepWDC can substantially enhance the accuracy of several baseline deterministic classification methods and outperforms state-of-the-art-methods on 2D and 3D data containing various types of perturbations (e.g., rotations, impulse noise, and downscaling).

[Finding All \\$\epsilon\\$-Good Arms in Stochastic Bandits](#)

- Blake Mason, Lalit Jain, Ardhendu Tripathy, Robert Nowak
- abstract@[open-review](#): The pure-exploration problem in stochastic multi-armed bandits aims to find one or more arms with the largest (or near largest) means. Examples include finding an ϵ -good arm, best-arm identification, top-\$k\$ arm identification, and finding all arms with means above a specified threshold. However, the problem of finding all ϵ -good arms has been overlooked in past work, although arguably this may be the most natural objective in many applications. For example, a virologist may conduct preliminary laboratory experiments on a large candidate set of treatments and move all ϵ -good treatments into more expensive clinical trials. Since the ultimate clinical efficacy is uncertain, it is important to identify all ϵ -good candidates. Mathematically, the all-\$\epsilon\$-good arm identification problem presents significant new challenges and surprises that do not arise in the pure-exploration objectives studied in the past. We introduce two algorithms to overcome these and demonstrate their great empirical performance on a large-scale crowd-sourced dataset of \$2.2M ratings collected by the New Yorker Caption Contest as well as a dataset testing hundreds of possible cancer drugs.

[Meta-Learning through Hebbian Plasticity in Random Networks](#)

- Elias Najarro, Sebastian Risi
- abstract@[open-review](#): Lifelong learning and adaptability are two defining aspects of biological agents. Modern reinforcement learning (RL) approaches have shown significant progress in solving complex tasks, however once training is concluded, the found solutions are typically static and incapable of adapting to new information or perturbations. While it is still not completely understood how biological brains learn and adapt so efficiently from experience, it is believed that synaptic plasticity plays a prominent role in this process. Inspired by this biological mechanism, we propose a search method that, instead of optimizing the weight parameters of neural networks directly, only searches for synapse-specific Hebbian learning rules that allow the network to continuously self-organize its weights during the lifetime of the agent. We demonstrate our approach on several reinforcement learning tasks with different sensory modalities and more than 450K trainable plasticity parameters. We find that starting from completely random weights, the discovered Hebbian rules enable an agent to navigate a dynamical 2D-pixel environment; likewise they allow a simulated 3D quadrupedal robot to learn how to walk while adapting to morphological damage not seen during training and in the absence of any explicit reward or error signal in less than 100 timesteps.

[A Computational Separation between Private Learning and Online Learning](#)

- Mark Bun
- abstract@[open-review](#): A recent line of work has shown a qualitative equivalence between differentially private PAC learning and online learning: A concept class is privately learnable if and only if it is online learnable with a finite mistake bound. However, both directions of this equivalence incur significant losses in both sample and computational efficiency. Studying a special case of this connection, Gonen, Hazan, and Moran (NeurIPS 2019) showed that uniform or highly sample-efficient pure-private learners can be time-efficiently compiled into online learners. We show that, assuming the existence of one-way functions, such an efficient conversion is impossible even for general pure-private learners with polynomial sample complexity. This resolves a question of Neel, Roth, and Wu (FOCS 2019).

[Top-KAST: Top-K Always Sparse Training](#)

- Siddhant Jayakumar, Razvan Pascanu, Jack Rae, Simon Osindero, Erich Elsen
- abstract@[open-review](#): Sparse neural networks are becoming increasingly important as the field seeks to improve the performance of existing models by scaling them up, while simultaneously trying to reduce power consumption and computational footprint. Unfortunately, most existing methods for inducing performant sparse models still entail the instantiation of dense parameters, or dense gradients in the backward-pass, during training. For very large models this requirement can be prohibitive. In this work we propose Top-KAST, a method that preserves constant sparsity throughout training (in both the forward and backward-passes). We demonstrate the efficacy of our approach by showing that it performs comparably to or better than previous works when training models on the established ImageNet benchmark, whilst fully maintaining sparsity. In addition to our ImageNet results, we also demonstrate our approach in the domain of language modeling where the current best performing architectures tend to have tens of billions of parameters and scaling up does not yet seem to have saturated performance. Sparse versions of these architectures can be run with significantly fewer resources, making them more widely accessible and applicable. Furthermore, in addition to being effective, our approach is straightforward and can easily be implemented in a wide range of existing machine learning frameworks with only a few additional lines of code. We therefore hope that our contribution will help enable the broader community to explore the potential held by massive models, without incurring massive computational cost.

[Meta-Learning with Adaptive Hyperparameters](#)

- Sungyong Baik, Myungsub Choi, Janghoon Choi, Heewon Kim, Kyoung Mu Lee
- abstract@[open-review](#): Despite its popularity, several recent works question the effectiveness of MAML when test tasks are different from training tasks, thus suggesting various task-conditioned methodology to improve the initialization. Instead of searching for better task-aware initialization, we focus on a complementary factor in MAML framework, inner-loop optimization (or fast adaptation). Consequently, we propose a new weight update rule that greatly enhances the fast adaptation process. Specifically, we introduce a small meta-network that can adaptively generate per-step hyperparameters: learning rate and weight decay coefficients. The experimental results validate that the Adaptive Learning of hyperparameters for Fast Adaptation (ALFA) is the equally important ingredient that was often neglected in the recent few-shot learning approaches. Surprisingly, fast adaptation from random initialization with ALFA can already outperform MAML.

[Tight last-iterate convergence rates for no-regret learning in multi-player games](#)

- Noah Golowich, Sarath Pattathil, Constantinos Daskalakis
- abstract@[open-review](#): We study the question of obtaining last-iterate convergence rates for no-regret learning algorithms in multi-player games. We show that the optimistic gradient (OG) algorithm with a constant step-size, which is no-regret, achieves a last-iterate rate of $O(1/\sqrt{T})$ with respect to the gap function in smooth monotone games. This result addresses a question of Mertikopoulos & Zhou (2018), who asked whether extra-gradient approaches (such as OG) can be applied to achieve improved guarantees in the multi-agent learning setting. The proof of our upper bound uses a new technique centered around an adaptive choice of potential function at each iteration. We also show that the $O(1/\sqrt{T})$ rate is tight for all p-SCLI algorithms, which includes OG as a special case. As a byproduct of our lower bound analysis we additionally present a proof of a conjecture of Arjevani et al. (2015) which is more direct than previous approaches.

[Curvature Regularization to Prevent Distortion in Graph Embedding](#)

- Hongbin Pei, Bingzhe Wei, Kevin Chang, Chunxu Zhang, Bo Yang
- abstract@[open-review](#): Recent research on graph embedding has achieved success in various applications. Most graph embedding methods preserve the proximity in a graph into a manifold in an embedding space. We argue an important but neglected problem about this proximity-preserving strategy: Graph topology patterns, while preserved well into an embedding manifold by preserving proximity, may distort in the ambient embedding Euclidean space, and hence to detect them becomes difficult for machine learning models. To address the problem, we propose curvature regularization, to enforce flatness for embedding manifolds, thereby preventing the distortion. We present a novel angle-based sectional curvature, termed ABS curvature, and accordingly three kinds of curvature regularization to induce flat embedding manifolds during graph embedding. We integrate curvature regularization into five popular proximity-preserving embedding methods, and empirical results in two applications show significant improvements on a wide range of open graph datasets.

[Perturbing Across the Feature Hierarchy to Improve Standard and Strict Blackbox Attack Transferability](#)

- Nathan Inkawich, Kevin Liang, Binghui Wang, Matthew Inkawich, Lawrence Carin, Yiran Chen
- abstract@[open-review](#): We consider the blackbox transfer-based targeted adversarial attack threat model in the realm of deep neural network (DNN) image classifiers. Rather than focusing on crossing decision boundaries at the output layer of the source model, our method perturbs representations throughout the extracted feature hierarchy to resemble other classes. We design a flexible attack framework that allows for multi-layer perturbations and demonstrates state-of-the-art targeted transfer performance between ImageNet DNNs. We also show the superiority of our feature space methods under a relaxation of the common assumption that the source and target models are trained on the same dataset and label space, in some instances achieving a \$10times\$ increase in targeted success rate relative to other blackbox transfer methods. Finally, we analyze why the proposed methods outperform existing attack strategies and show an extension of the method in the case when limited queries to the blackbox model are allowed.

Statistical and Topological Properties of Sliced Probability Divergences

- Kimia Nadjahi, Alain Durmus, LÃ©naÃ©c Chizat, Soheil Kolouri, Shahin Shahrampour, Umut Simsekli
- abstract@[open-review](#): The idea of slicing divergences has been proven to be successful when comparing two probability measures in various machine learning applications including generative modeling, and consists in computing the expected value of a 'base divergence' between \emph{one-dimensional random projections} of the two measures. However, the topological, statistical, and computational consequences of this technique have not yet been well-established. In this paper, we aim at bridging this gap and derive various theoretical properties of sliced probability divergences. First, we show that slicing preserves the metric axioms and the weak continuity of the divergence, implying that the sliced divergence will share similar topological properties. We then precise the results in the case where the base divergence belongs to the class of integral probability metrics. On the other hand, we establish that, under mild conditions, the sample complexity of a sliced divergence does not depend on the problem dimension. We finally apply our general results to several base divergences, and illustrate our theory on both synthetic and real data experiments.

Probabilistic Active Meta-Learning

- Jean Kaddour, Steindor Saemundsson, Marc Deisenroth (he/him)
- abstract@[open-review](#): Data-efficient learning algorithms are essential in many practical applications where data collection is expensive, e.g., in robotics due to the wear and tear. To address this problem, meta-learning algorithms use prior experience about tasks to learn new, related tasks efficiently. Typically, a set of training tasks is assumed given or randomly chosen. However, this setting does not take into account the sequential nature that naturally arises when training a model from scratch in real-life: how do we collect a set of training tasks in a data-efficient manner? In this work, we introduce task selection based on prior experience into a meta-learning algorithm by conceptualizing the learner and the active meta-learning setting using a probabilistic latent variable model. We provide empirical evidence that our approach improves data-efficiency when compared to strong baselines on simulated robotic experiments.

Knowledge Distillation in Wide Neural Networks: Risk Bound, Data Efficiency and Imperfect Teacher

- Guangda Ji, Zhanxing Zhu
- abstract@[open-review](#): Knowledge distillation is a strategy of training a student network with guide of the soft output from a teacher network. It has been a successful method of model compression and knowledge transfer. However, currently knowledge distillation lacks a convincing theoretical understanding. On the other hand, recent finding on neural tangent kernel enables us to approximate a wide neural network with a linear model of the network's random features. In this paper, we theoretically analyze the knowledge distillation of a wide neural network. First we provide a transfer risk bound for the linearized model of the network. Then we propose a metric of the task's training difficulty, called data inefficiency. Based on this metric, we show that for a perfect teacher, a high ratio of teacher's soft labels can be beneficial. Finally, for the case of imperfect teacher, we find that hard labels can correct teacher's wrong prediction, which explains the practice of mixing hard and soft labels.

Adversarial Attacks on Deep Graph Matching

- Zijie Zhang, Zeru Zhang, Yang Zhou, Yelong Shen, Ruoming Jin, Dejing Dou
- abstract@[open-review](#): Despite achieving remarkable performance, deep graph learning models, such as node classification and network embedding, suffer from harassment caused by small adversarial perturbations. However, the vulnerability analysis of graph matching under adversarial attacks has not been fully investigated yet. This paper proposes an adversarial attack model with two novel attack techniques to perturb the graph structure and degrade the quality of deep graph matching: (1) a kernel density estimation approach is utilized to estimate and maximize node densities to derive imperceptible perturbations, by pushing attacked nodes to dense regions in two graphs, such that they are indistinguishable from many neighbors; and (2) a meta learning-based projected gradient descent method is developed to well choose attack starting points and to improve the search performance for producing effective perturbations. We evaluate the effectiveness of the attack model on real datasets and validate that the attacks can be transferable to other graph learning models.

The Generalization-Stability Tradeoff In Neural Network Pruning

- Brian Bartoldson, Ari Morcos, Adrian Barbu, Gordon Erlebacher
- abstract@[open-review](#): Pruning neural network parameters is often viewed as a means to compress models, but pruning has also been motivated by the desire to prevent overfitting. This motivation is particularly relevant given the perhaps surprising observation that a wide variety of pruning approaches increase test accuracy despite sometimes massive reductions in parameter counts. To better understand this phenomenon, we analyze the behavior of pruning over the course of training, finding that pruning's benefit to generalization increases with pruning's instability (defined as the drop in test accuracy immediately following pruning). We demonstrate that this "generalization-stability tradeoff" is present across a wide variety of pruning settings and propose a mechanism for its cause: pruning regularizes similarly to noise injection. Supporting this, we find less pruning stability leads to more model flatness and the benefits of pruning do not depend on permanent parameter removal. These results explain the compatibility of pruning-based generalization improvements and the high generalization recently observed in overparameterized networks.

Gradient-EM Bayesian Meta-Learning

- Yayi Zou, Xiaoqi Lu
- abstract@[open-review](#): Bayesian meta-learning enables robust and fast adaptation to new tasks with uncertainty assessment. The key idea behind Bayesian meta-learning is empirical Bayes inference of hierarchical model. In this work, we extend this framework to include a variety of existing methods, before proposing our variant based on gradient-EM algorithm. Our method improves computational efficiency by avoiding back-propagation computation in the meta-update step, which is exhausting for deep neural networks. Furthermore, it provides flexibility to the inner-update optimization procedure by decoupling it from meta-update. Experiments on sinusoidal regression, few-shot image classification, and policy-based reinforcement learning show that our method not only achieves better accuracy with less computation cost, but is also more robust to uncertainty.

Logarithmic Regret Bound in Partially Observable Linear Dynamical Systems

- Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, Anima Anandkumar
- abstract@[open-review](#): We study the problem of system identification and adaptive control in partially observable linear dynamical systems. Adaptive and closed-loop system identification is a challenging problem due to correlations introduced in data collection. In this paper, we present the first model estimation method with finite-time guarantees in both open and closed-loop system identification. Deploying this estimation method, we propose adaptive control online learning (AdapOn), an efficient reinforcement learning algorithm that adaptively learns the system dynamics and continuously updates its controller through online learning steps. AdapOn estimates the model dynamics by occasionally solving a linear regression problem through interactions with the environment. Using policy re-parameterization and the estimated model, AdapOn constructs counterfactual loss functions to be used for updating the controller through online gradient descent. Over time, AdapOn improves its model estimates and obtains more accurate gradient updates to improve the controller. We show that AdapOn achieves a regret upper bound of $\text{polylog}(\text{left}(T\text{right}))$, after T time steps of agent-environment interaction. To the best of our knowledge, AdapOn is the first algorithm that achieves $\text{polylog}(\text{left}(T\text{right}))$ regret in adaptive control of unknown partially observable linear dynamical systems which includes linear quadratic Gaussian (LQG) control.

Linearly Converging Error Compensated SGD

- Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, Peter Richtarik
- abstract@[open-review](#): In this paper, we propose a unified analysis of variants of distributed SGD with arbitrary compressions and delayed updates. Our framework is general enough to cover different variants of quantized SGD, Error-Compensated SGD (EC-SGD), and SGD with delayed updates (D-SGD). Via single theorem, we derive the complexity results for all the methods that fit our framework. For the existing methods, this theorem gives the best-known complexity results. Moreover, using our general scheme, we develop new variants of SGD that combine variance reduction or arbitrary sampling with error feedback and quantization and derive the convergence rates for these methods beating the state-of-the-art results. In order to illustrate the strength of our framework, we develop 16 new methods that fit this. In particular, we propose the first method called EC-SGD-DIANA that is based on error-feedback for biased compression operator and quantization of gradient differences and prove the convergence guarantees showing that EC-SGD-DIANA converges to the exact optimum asymptotically in expectation with constant learning rate for both convex and strongly convex objectives when workers compute full gradients of their loss functions. Moreover, for the case when the loss function of the worker has the form of finite sum, we modified the method and got a new one called EC-LSVRG-DIANA which is the first distributed stochastic method with error feedback and variance reduction that converges to the exact optimum asymptotically in expectation with constant learning rate.

Canonical 3D Deformer Maps: Unifying parametric and non-parametric methods for dense weakly-supervised category reconstruction

- David Novotny, Roman Shapovalov, Andrea Vedaldi
- abstract@[open-review](#): We propose the Canonical 3D Deformer Map, a new representation of the 3D shape of common object categories that can be learned from a collection of 2D images of independent objects. Our method builds in a novel way on concepts from parametric deformation models, non-parametric 3D reconstruction, and canonical embeddings, combining their individual advantages. In particular, it learns to associate each image pixel with a deformation model of the corresponding 3D object point which is canonical, i.e. intrinsic to the identity of the point and shared across objects of the category. The result is a method that, given only sparse 2D supervision at training time, can, at test time, reconstruct the 3D shape and texture of objects from single views, while establishing meaningful dense correspondences between object instances. It also achieves state-of-the-art results in dense 3D reconstruction on public in-the-wild datasets of faces, cars, and birds.

A Self-Tuning Actor-Critic Algorithm

- Tom Zahavy, Zhongwen Xu, Vivek Veeriah, Matteo Hessel, Junhyuk Oh, Hado P. van Hasselt, David Silver, Satinder Singh
- abstract@[open-review](#): Reinforcement learning algorithms are highly sensitive to the choice of hyperparameters, typically requiring significant manual effort to identify hyperparameters that perform well on a new domain. In this paper, we take a step towards addressing this issue by using metagradients to automatically adapt hyperparameters online by meta-gradient descent (Xu et al., 2018). We apply our algorithm, Self-Tuning Actor-Critic (STAC), to self-tune all the differentiable hyperparameters of an actor-critic loss function, to discover auxiliary tasks, and to improve off-policy learning using a novel leaky V-trace operator. STAC is simple to use, sample efficient and does not require a significant increase in compute. Ablative studies show that the overall performance of STAC improved as we adapt more hyperparameters. When applied to the Arcade Learning Environment (Bellemare et al. 2012), STAC improved the median human normalized score in 200M steps from 243% to 364%. When applied to the DM Control suite (Tassa et al., 2018), STAC improved the mean score in 30M steps from 217 to 389 when learning with features, from 108 to 202 when learning from pixels, and from 195 to 295 in the Real-World Reinforcement Learning Challenge (Dulac-Arnold et al., 2020).

The Cone of Silence: Speech Separation by Localization

- Teerapat Jenrungrot, Vivek Jayaram, Steve Seitz, Ira Kemelmacher-Shlizerman
- abstract@[open-review](#): Given a multi-microphone recording of an unknown number of speakers talking concurrently, we simultaneously localize the sources and separate the individual speakers. At the core of our method is a deep network, in the waveform domain, which isolates sources within an angular region $\pm w/2$, given an angle of interest θ and angular window size w . By exponentially decreasing w , we can perform a binary search to localize and separate all sources in logarithmic time. Our algorithm also allows for an arbitrary number of potentially moving speakers at test time, including more speakers than seen during training. Experiments demonstrate state of the art performance for both source separation and source localization, particularly in high levels of background noise.

High-Dimensional Bayesian Optimization via Nested Riemannian Manifolds

- Noémie Jaquier, Leonel Rozo
- abstract@[open-review](#): Despite the recent success of Bayesian optimization (BO) in a variety of applications where sample efficiency is imperative, its performance may be seriously compromised in settings characterized by high-dimensional parameter spaces. A solution to preserve the sample efficiency of BO in such problems is to introduce domain knowledge into its formulation. In this paper, we propose to exploit the geometry of non-Euclidean search spaces, which often arise in a variety of domains, to learn structure-preserving mappings and optimize the acquisition function of BO in low-dimensional latent spaces. Our approach, built on Riemannian manifolds theory, features geometry-aware Gaussian processes that jointly learn a nested-manifolds embedding and a representation of the objective function in the latent space. We test our approach in several benchmark artificial landscapes and report that it not only outperforms other high-dimensional BO approaches in several settings, but consistently optimizes the objective functions, as opposed to geometry-unaware BO methods.

Train-by-Reconnect: Decoupling Locations of Weights from Their Values

- Yushi Qiu, Reiji Suda
- abstract@[open-review](#): What makes untrained deep neural networks (DNNs) different from the trained performant ones? By zooming into the weights in well-trained DNNs, we found that it is the location of weights that holds most of the information encoded by the training. Motivated by this observation, we hypothesized that weights in DNNs trained using stochastic gradient-based methods can be separated into two dimensions: the location of weights, and their exact values. To assess our hypothesis, we propose a novel method called lookahead permutation (LaPerm) to train DNNs by reconnecting the weights. We empirically demonstrate LaPerm's versatility while producing extensive evidence to support our hypothesis: when the initial weights are random and dense, our method demonstrates speed and performance similar to or better than that of regular optimizers, e.g., Adam. When the initial weights are random and sparse (many zeros), our method changes the way neurons connect, achieving accuracy comparable to that of a well-trained dense network. When the initial weights share a single value, our method finds a weight agnostic neural network with far-better-than-chance accuracy.

Learning discrete distributions: user vs item-level privacy

- Yuhan Liu, Ananda Theertha Suresh, Felix Xinnan X. Yu, Sanjiv Kumar, Michael Riley
- abstract@[open-review](#): Much of the literature on differential privacy focuses on item-level privacy, where loosely speaking, the goal is to provide privacy per item or training example. However, recently many practical applications such as federated learning require preserving privacy for all items of a single user, which is much harder to achieve. Therefore understanding the theoretical limit of user-level privacy becomes crucial.

We study the fundamental problem of learning discrete distributions over $\$k\$$ symbols with user-level differential privacy. If each user has $\$m\$$ samples, we show that straightforward applications of Laplace or Gaussian mechanisms require the number of users to be $\$O(k/(m\alpha^2) + k/\epsilon\alpha)\$$ to achieve an $\|\ell_1$ distance of $\|\alpha\|$ between the true and estimated distributions, with the privacy-induced penalty $\$k/\epsilon\alpha\$$ independent of the number of samples per user $\$m\$$. Moreover, we show that any mechanism that only operates on the final aggregate should require a user complexity of the same order. We then propose a mechanism such that the number of users scales as $\|\tilde{O}(k/(m\alpha^2) + k\sqrt{m}/\epsilon\alpha)\|$ and further show that it is nearly-optimal under certain regimes. Thus the privacy penalty is $\|\tilde{\Theta}(\sqrt{m})\|$ times smaller compared to the standard mechanisms.

We also propose general techniques for obtaining lower bounds on restricted differentially private estimators and a lower bound on the total variation between binomial distributions, both of which might be of independent interest.

[Matrix Completion with Quantified Uncertainty through Low Rank Gaussian Copula](#)

- Yuxuan Zhao, Madeleine Udell
- abstract@[open-review](#): Modern large scale datasets are often plagued with missing entries. For tabular data with missing values, a flurry of imputation algorithms solve for a complete matrix which minimizes some penalized reconstruction error. However, almost none of them can estimate the uncertainty of its imputations. This paper proposes a probabilistic and scalable framework for missing value imputation with quantified uncertainty. Our model, the Low Rank Gaussian Copula, augments a standard probabilistic model, Probabilistic Principal Component Analysis, with marginal transformations for each column that allow the model to better match the distribution of the data. It naturally handles Boolean, ordinal, and real-valued observations and quantifies the uncertainty in each imputation. The time required to fit the model scales linearly with the number of rows and the number of columns in the dataset. Empirical results show the method yields state-of-the-art imputation accuracy across a wide range of data types, including those with high rank. Our uncertainty measure predicts imputation error well: entries with lower uncertainty do have lower imputation error (on average). Moreover, for real-valued data, the resulting confidence intervals are well-calibrated.

[Sparse and Continuous Attention Mechanisms](#)

- AndrÃ© Martins, AntÃ³nio Farinhas, Marcos Treviso, Vlad Niculae, Pedro Aguiar, Mario Figueiredo
- abstract@[open-review](#): Exponential families are widely used in machine learning; they include many distributions in continuous and discrete domains (e.g., Gaussian, Dirichlet, Poisson, and categorical distributions via the softmax transformation). Distributions in each of these families have fixed support. In contrast, for finite domains, there has been recent work on sparse alternatives to softmax (e.g., sparsemax and alpha-entmax), which have varying support, being able to assign zero probability to irrelevant categories. These discrete sparse mappings have been used for improving interpretability of neural attention mechanisms. This paper expands that work in two directions: first, we extend alpha-entmax to continuous domains, revealing a link with Tsallis statistics and deformed exponential families. Second, we introduce continuous-domain attention mechanisms, deriving efficient gradient backpropagation algorithms for alpha in {1,2}. Experiments on attention-based text classification, machine translation, and visual question answering illustrate the use of continuous attention in 1D and 2D, showing that it allows attending to time intervals and compact regions.

[Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection](#)

- Xiang Li, Wenhui Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, Jian Yang
- abstract@[open-review](#): One-stage detector basically formulates object detection as dense classification and localization (i.e., bounding box regression). The classification is usually optimized by Focal Loss and the box location is commonly learned under Dirac delta distribution. A recent trend for one-stage detectors is to introduce an \emph{individual} prediction branch to estimate the quality of localization, where the predicted quality facilitates the classification to improve detection performance. This paper delves into the \emph{representations} of the above three fundamental elements: quality estimation, classification and localization. Two problems are discovered in existing practices, including (1) the inconsistent usage of the quality estimation and classification between training and inference, and (2) the inflexible Dirac delta distribution for localization. To address the problems, we design new representations for these elements. Specifically, we merge the quality estimation into the class prediction vector to form a joint representation, and use a vector to represent arbitrary distribution of box locations. The improved representations eliminate the inconsistency risk and accurately depict the flexible distribution in real data, but contain \emph{continuous} labels, which is beyond the scope of Focal Loss. We then propose Generalized Focal Loss (GFL) that generalizes Focal Loss from its discrete form to the \emph{continuous} version for successful optimization. On COCO \{tt test-dev\}, GFL achieves 45.0\% AP using ResNet-101 backbone, surpassing state-of-the-art SAPD (43.5\%) and ATSS (43.6\%) with higher or comparable inference speed.

[Learning by Minimizing the Sum of Ranked Range](#)

- Shu Hu, Yiming Ying, xin wang, Siwei Lyu
- abstract@[open-review](#): In forming learning objectives, one oftentimes needs to aggregate a set of individual values to a single output. Such cases occur in the aggregate loss, which combines individual losses of a learning model over each training sample, and in the individual loss for multi-label learning, which combines prediction scores over all class labels. In this work, we introduce the sum of ranked range (SoRR) as a general approach to form learning objectives. A ranked range is a consecutive sequence of sorted values of a set of real numbers. The minimization of SoRR is solved with the difference of convex algorithm (DCA). We explore two applications in machine learning of the minimization of the SoRR framework, namely the AoRR aggregate loss for binary classification and the TKML individual loss for multi-label/multi-class classification. Our empirical results highlight the effectiveness of the proposed optimization framework and demonstrate the applicability of proposed losses using synthetic and real datasets.

[Robust Deep Reinforcement Learning against Adversarial Perturbations on State Observations](#)

- Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, Cho-Jui Hsieh
- abstract@[open-review](#): A deep reinforcement learning (DRL) agent observes its states through observations, which may contain natural measurement errors or adversarial noises. Since the observations deviate from the true states, they can mislead the agent into making suboptimal actions. Several works have shown this vulnerability via adversarial attacks, but how to improve the robustness of DRL under this setting has not been well studied. We show that naively applying existing techniques on improving robustness for classification tasks, like adversarial training, are ineffective for many RL tasks. We propose the state-adversarial Markov decision process (SA-MDP) to study the fundamental properties of this problem, and develop a theoretically principled policy regularization which can be applied to a large family of DRL algorithms, including deep deterministic policy gradient (DDPG), proximal policy optimization (PPO) and deep Q networks (DQN), for both discrete and continuous action control problems. We significantly improve the robustness of DDPG, PPO and DQN agents under a suite of strong white box adversarial attacks, including two new attacks of our own. Additionally, we find that a robust policy noticeably improves DRL performance in a number of environments.

[Understanding Anomaly Detection with Deep Invertible Networks through Hierarchies of Distributions and Features](#)

- Robin Schirrmeyer, Yuxuan Zhou, Tonio Ball, Dan Zhang
- abstract@[open-review](#): Deep generative networks trained via maximum likelihood on a natural image dataset like CIFAR10 often assign high likelihoods to images from datasets with different objects (e.g., SVHN). We refine previous investigations of this failure at anomaly detection for invertible generative networks and provide a clear explanation of it as a combination of model bias and domain prior: Convolutional networks learn similar low-level feature distributions when trained on any natural image dataset and these low-level features dominate the likelihood. Hence, when the discriminative features between inliers and outliers are on a high-level, e.g., object shapes, anomaly detection becomes particularly challenging. To remove the negative impact of

model bias and domain prior on detecting high-level differences, we propose two methods, first, using the log likelihood ratios of two identical models, one trained on the in-distribution data (e.g., CIFAR10) and the other one on a more general distribution of images (e.g., 80 Million Tiny Images). We also derive a novel outlier loss for the in-distribution network on samples from the more general distribution to further improve the performance. Secondly, using a multi-scale model like Glow, we show that low-level features are mainly captured at early scales. Therefore, using only the likelihood contribution of the final scale performs remarkably well for detecting high-level feature differences of the out-of-distribution and the in-distribution. This method is especially useful if one does not have access to a suitable general distribution. Overall, our methods achieve strong anomaly detection performance in the unsupervised setting, and only slightly underperform state-of-the-art classifier-based methods in the supervised setting. Code can be found at <https://github.com/boschresearch/hierarchicalanomalydetection>.

[Fair Hierarchical Clustering](#)

- Sara Ahmadian, Alessandro Epasto, Marina Knittel, Ravi Kumar, Mohammad Mahdian, Benjamin Moseley, Philip Pham, Sergei Vassilvitskii, Yuyan Wang
- abstract@[open-review](#): As machine learning has become more prevalent, researchers have begun to recognize the necessity of ensuring machine learning systems are fair. Recently, there has been an interest in defining a notion of fairness that mitigates over-representation in traditional clustering. In this paper we extend this notion to hierarchical clustering, where the goal is to recursively partition the data to optimize a specific objective. For various natural objectives, we obtain simple, efficient algorithms to find a provably good fair hierarchical clustering. Empirically, we show that our algorithms can find a fair hierarchical clustering, with only a negligible loss in the objective.

[Self-training Avoids Using Spurious Features Under Domain Shift](#)

- Yining Chen, Colin Wei, Ananya Kumar, Tengyu Ma
- abstract@[open-review](#): In unsupervised domain adaptation, existing theory focuses on situations where the source and target domains are close. In practice, conditional entropy minimization and pseudo-labeling work even when the domain shifts are much larger than those analyzed by existing theory. We identify and analyze one particular setting where the domain shift can be large, but these algorithms provably work: certain spurious features correlate with the label in the source domain but are independent of the label in the target. Our analysis considers linear classification where the spurious features are Gaussian and the non-spurious features are a mixture of log-concave distributions. For this setting, we prove that entropy minimization on unlabeled target data will avoid using the spurious feature if initialized with a decently accurate source classifier, even though the objective is non-convex and contains multiple bad local minima using the spurious features. We verify our theory for spurious domain shift tasks on semi-synthetic Celeb-A and MNIST datasets. Our results suggest that practitioners collect and self-train on large, diverse datasets to reduce biases in classifiers even if labeling is impractical.

[Improving Online Rent-or-Buy Algorithms with Sequential Decision Making and ML Predictions](#)

- Shom Banerjee
- abstract@[open-review](#): In this work we study online rent-or-buy problems as a sequential decision making problem. We show how one can integrate predictions, typically coming from a machine learning (ML) setup, into this framework. Specifically, we consider the ski-rental problem and the dynamic TCP acknowledgment problem. We present new online algorithms and obtain explicit performance bounds in-terms of the accuracy of the prediction. Our algorithms are close to optimal with accurate predictions while hedging against less accurate predictions.

[CircleGAN: Generative Adversarial Learning across Spherical Circles](#)

- Woohyeon Shim, Minsu Cho
- abstract@[open-review](#): We present a novel discriminator for GANs that improves realness and diversity of generated samples by learning a structured hypersphere embedding space using spherical circles. The proposed discriminator learns to populate realistic samples around the longest spherical circle, i.e., a great circle, while pushing unrealistic samples toward the poles perpendicular to the great circle. Since longer circles occupy larger area on the hypersphere, they encourage more diversity in representation learning, and vice versa. Discriminating samples based on their corresponding spherical circles can thus naturally induce diversity to generated samples. We also extend the proposed method for conditional settings with class labels by creating a hypersphere for each category and performing class-wise discrimination and update. In experiments, we validate the effectiveness for both unconditional and conditional generation on standard benchmarks, achieving the state of the art.

[WOR and \\$p\\$'s: Sketches for \\$\ell_p\\$-Sampling Without Replacement](#)

- Edith Cohen, Rasmus Pagh, David Woodruff
- abstract@[open-review](#): Weighted sampling is a fundamental tool in data analysis and machine learning pipelines. Samples are used for efficient estimation of statistics or as sparse representations of the data. When weight distributions are skewed, as is often the case in practice, without-replacement (WOR) sampling is much more effective than with-replacement (WR) sampling: It provides a broader representation and higher accuracy for the same number of samples. We design novel composable sketches for WOR $\{\text{em} \ell_p\text{-sampling}\}$, weighted sampling of keys according to a power $p \in [0, 2]$ of their frequency (or for signed data, sum of updates). Our sketches have size that grows only linearly with sample size. Our design is simple and practical, despite intricate analysis, and based on off-the-shelf use of widely implemented heavy hitters sketches such as CountSketch . Our method is the first to provide WOR sampling in the important regime of $p > 1$ and the first to handle signed updates for $p > 0$.

[Hypersolvers: Toward Fast Continuous-Depth Models](#)

- Michael Poli, Stefano Massaroli, Atsushi Yamashita, Hajime Asama, Jinkyoo Park
- abstract@[open-review](#): The infinite-depth paradigm pioneered by Neural ODEs has launched a renaissance in the search for novel dynamical system-inspired deep learning primitives; however, their utilization in problems of non-trivial size has often proved impossible due to poor computational scalability. This work paves the way for scalable Neural ODEs with time-to-prediction comparable to traditional discrete networks. We introduce hypersolvers, neural networks designed to solve ODEs with low overhead and theoretical guarantees on accuracy. The synergistic combination of hypersolvers and Neural ODEs allows for cheap inference and unlocks a new frontier for practical application of continuous-depth models. Experimental evaluations on standard benchmarks, such as sampling for continuous normalizing flows, reveal consistent pareto efficiency over classical numerical methods.

[Log-Likelihood Ratio Minimizing Flows: Towards Robust and Quantifiable Neural Distribution Alignment](#)

- Ben Usman, Avneesh Sud, Nick Dufour, Kate Saenko
- abstract@[open-review](#): Distribution alignment has many applications in deep learning, including domain adaptation and unsupervised image-to-image translation. Most prior work on unsupervised distribution alignment relies either on minimizing simple non-parametric statistical distances such as maximum mean discrepancy or on adversarial alignment. However, the former fails to capture the structure of complex real-world distributions, while the latter is difficult to train and does not provide any universal convergence guarantees or automatic quantitative validation procedures. In this paper, we propose a new distribution alignment method based on a log-likelihood ratio statistic and normalizing flows. We show that, under certain assumptions, this

combination yields a deep neural likelihood-based minimization objective that attains a known lower bound upon convergence. We experimentally verify that minimizing the resulting objective results in domain alignment that preserves the local structure of input domains.

[Escaping the Gravitational Pull of Softmax](#)

- Jincheng Mei, Chenjun Xiao, Bo Dai, Lihong Li, Csaba Szepesvari, Dale Schuurmans
- abstract@[open-review](#): The softmax is the standard transformation used in machine learning to map real-valued vectors to categorical distributions. Unfortunately, this transform poses serious drawbacks for gradient descent (ascent) optimization. We reveal this difficulty by establishing two negative results: (1) optimizing any expectation with respect to the softmax must exhibit sensitivity to parameter initialization ("softmax gravity well"), and (2) optimizing log-probabilities under the softmax must exhibit slow convergence ("softmax damping"). Both findings are based on an analysis of convergence rates using the Non-uniform $\mathcal{L}\{\cdot\}$ -ojasiewicz ($\mathcal{NL}\{\cdot\}$) inequalities. To circumvent these shortcomings we investigate an alternative transformation, the escort mapping, that demonstrates better optimization properties. The disadvantages of the softmax and the effectiveness of the escort transformation are further explained using the concept of $\mathcal{NL}\{\cdot\}$ coefficient. In addition to proving bounds on convergence rates to firmly establish these results, we also provide experimental evidence for the superiority of the escort transformation.

[Regret in Online Recommendation Systems](#)

- Kaito Ariu, Narae Ryu, Se-Young Yun, Alexandre Proutiere
- abstract@[open-review](#): This paper proposes a theoretical analysis of recommendation systems in an online setting, where items are sequentially recommended to users over time. In each round, a user, randomly picked from a population of m users, arrives. The decision-maker observes the user and selects an item from a catalogue of n items. Importantly, an item cannot be recommended twice to the same user. The probabilities that a user likes each item are unknown, and the performance of the recommendation algorithm is captured through its regret, considering as a reference an Oracle algorithm aware of these probabilities. We investigate various structural assumptions on these probabilities: we derive for each of them regret lower bounds, and devise algorithms achieving these limits. Interestingly, our analysis reveals the relative weights of the different components of regret: the component due to the constraint of not presenting the same item twice to the same user, that due to learning the chances users like items, and finally that arising when learning the underlying structure.

[On Convergence and Generalization of Dropout Training](#)

- Poorya Mianjy, Raman Arora
- abstract@[open-review](#): We study dropout in two-layer neural networks with rectified linear unit (ReLU) activations. Under mild overparametrization and assuming that the limiting kernel can separate the data distribution with a positive margin, we show that the dropout training with logistic loss achieves ϵ -suboptimality in the test error in $O(1/\epsilon)$ iterations.

[Second Order Optimality in Decentralized Non-Convex Optimization via Perturbed Gradient Tracking](#)

- Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari
- abstract@[open-review](#): In this paper we study the problem of escaping from saddle points and achieving second-order optimality in a decentralized setting where a group of agents collaborate to minimize their aggregate objective function. We provide a non-asymptotic (finite-time) analysis and show that by following the idea of perturbed gradient descent, it is possible to converge to a second-order stationary point in a number of iterations which depends linearly on dimension and polynomially on the accuracy of second-order stationary point. Doing this in a communication-efficient manner requires overcoming several challenges, from identifying (first order) stationary points in a distributed manner, to adapting the perturbed gradient framework without prohibitive communication complexity. Our proposed Perturbed Decentralized Gradient Tracking (PDGT) method consists of two major stages: (i) a gradient-based step to find a first-order stationary point and (ii) a perturbed gradient descent step to escape from a first-order stationary point, if it is a saddle point with sufficient curvature. As a side benefit of our result, in the case that all saddle points are non-degenerate (strict), the proposed PDGT method finds a local minimum of the considered decentralized optimization problem in a finite number of iterations.

[Implicit Regularization in Deep Learning May Not Be Explainable by Norms](#)

- Noam Razin, Nadav Cohen
- abstract@[open-review](#): Mathematically characterizing the implicit regularization induced by gradient-based optimization is a longstanding pursuit in the theory of deep learning. A widespread hope is that a characterization based on minimization of norms may apply, and a standard test-bed for studying this prospect is matrix factorization (matrix completion via linear neural networks). It is an open question whether norms can explain the implicit regularization in matrix factorization. The current paper resolves this open question in the negative, by proving that there exist natural matrix factorization problems on which the implicit regularization drives all norms (and quasi-norms) towards infinity. Our results suggest that, rather than perceiving the implicit regularization via norms, a potentially more useful interpretation is minimization of rank. We demonstrate empirically that this interpretation extends to a certain class of non-linear neural networks, and hypothesize that it may be key to explaining generalization in deep learning.

[POMO: Policy Optimization with Multiple Optima for Reinforcement Learning](#)

- Yeong-Dae Kwon, Jinho Choo, Byoungjin Kim, Iljoo Yoon, Youngjune Gwon, Seungjai Min
- abstract@[open-review](#): In neural combinatorial optimization (CO), reinforcement learning (RL) can turn a deep neural net into a fast, powerful heuristic solver of NP-hard problems. This approach has a great potential in practical applications because it allows near-optimal solutions to be found without expert guides armed with substantial domain knowledge. We introduce Policy Optimization with Multiple Optima (POMO), an end-to-end approach for building such a heuristic solver. POMO is applicable to a wide range of CO problems. It is designed to exploit the symmetries in the representation of a CO solution. POMO uses a modified REINFORCE algorithm that forces diverse rollouts towards all optimal solutions. Empirically, the low-variance baseline of POMO makes RL training fast and stable, and it is more resistant to local minima compared to previous approaches. We also introduce a new augmentation-based inference method, which accompanies POMO nicely. We demonstrate the effectiveness of POMO by solving three popular NP-hard problems, namely, traveling salesman (TSP), capacitated vehicle routing (CVRP), and 0-1 knapsack (KP). For all three, our solver based on POMO shows a significant improvement in performance over all recent learned heuristics. In particular, we achieve the optimality gap of 0.14% with TSP100 while reducing inference time by more than an order of magnitude.

[Uncertainty-aware Self-training for Few-shot Text Classification](#)

- Subhabrata Mukherjee, Ahmed Awadallah
- abstract@[open-review](#): Recent success of pre-trained language models crucially hinges on fine-tuning them on large amounts of labeled data for the downstream task, that are typically expensive to acquire or difficult to access for many applications. We study self-training as one of the earliest semi-supervised learning approaches to reduce the annotation bottleneck by making use of large-scale unlabeled data for the target task. Standard self-training mechanism randomly samples instances from the unlabeled pool to generate pseudo-labels and augment labeled data. We propose an approach to improve self-training by incorporating uncertainty estimates of the underlying neural network leveraging recent advances in Bayesian deep learning. Specifically, we propose (i) acquisition functions to select instances from the unlabeled pool leveraging Monte Carlo (MC) Dropout, and (ii) learning mechanism

leveraging model confidence for self-training. As an application, we focus on text classification with five benchmark datasets. We show our methods leveraging only 20-30 labeled samples per class for each task for training and for validation perform within 3% of fully supervised pre-trained language models fine-tuned on thousands of labels with an aggregate accuracy of 91% and improvement of up to 12% over baselines.

[Learning to Learn with Feedback and Local Plasticity](#)

- Jack Lindsey, Ashok Litwin-Kumar
- abstract@[open-review](#): Interest in biologically inspired alternatives to backpropagation is driven by the desire to both advance connections between deep learning and neuroscience and address backpropagation's shortcomings on tasks such as online, continual learning. However, local synaptic learning rules like those employed by the brain have so far failed to match the performance of backpropagation in deep networks. In this study, we employ meta-learning to discover networks that learn using feedback connections and local, biologically inspired learning rules. Importantly, the feedback connections are not tied to the feedforward weights, avoiding biologically implausible weight transport. Our experiments show that meta-trained networks effectively use feedback connections to perform online credit assignment in multi-layer architectures. Surprisingly, this approach matches or exceeds a state-of-the-art gradient-based online meta-learning algorithm on regression and classification tasks, excelling in particular at continual learning. Analysis of the weight updates employed by these models reveals that they differ qualitatively from gradient descent in a way that reduces interference between updates. Our results suggest the existence of a class of biologically plausible learning mechanisms that not only match gradient descent-based learning, but also overcome its limitations.

[Every View Counts: Cross-View Consistency in 3D Object Detection with Hybrid-Cylindrical-Spherical Voxelization](#)

- Qi Chen, Lin Sun, Ernest Cheung, Alan L. Yuille
- abstract@[open-review](#): Recent voxel-based 3D object detectors for autonomous vehicles learn point cloud representations either from bird eye view (BEV) or range view (RV, a.k.a. the perspective view). However, each view has its own strengths and weaknesses. In this paper, we present a novel framework to unify and leverage the benefits from both BEV and RV. The widely-used cuboid-shaped voxels in Cartesian coordinate system only benefit learning BEV feature map. Therefore, to enable learning both BEV and RV feature maps, we introduce Hybrid-Cylindrical-Spherical voxelization. Our findings show that simply adding detection on another view as auxiliary supervision will lead to poor performance. We proposed a pair of cross-view transformers to transform the feature maps into the other view and introduce cross-view consistency loss on them. Comprehensive experiments on the challenging NuScenes Dataset validate the effectiveness of our proposed method by virtue of joint optimization and complementary information on both views. Remarkably, our approach achieved mAP of 55.8%, outperforming all published approaches by at least 3% in overall performance and up to 16.5% in safety-crucial categories like cyclist.

[Sharper Generalization Bounds for Pairwise Learning](#)

- Yunwen Lei, Antoine Ledent, Marius Kloft
- abstract@[open-review](#): Pairwise learning refers to learning tasks with loss functions depending on a pair of training examples, which includes ranking and metric learning as specific examples. Recently, there has been an increasing amount of attention on the generalization analysis of pairwise learning to understand its practical behavior. However, the existing stability analysis provides suboptimal high-probability generalization bounds. In this paper, we provide a refined stability analysis by developing generalization bounds which can be \sqrt{n} -times faster than the existing results, where n is the sample size. This implies excess risk bounds of the order $O(n^{-1/2})$ (up to a logarithmic factor) for both regularized risk minimization and stochastic gradient descent. We also introduce a new on-average stability measure to develop optimistic bounds in a low noise setting. We apply our results to ranking and metric learning, and clearly show the advantage of our generalization bounds over the existing analysis.

[A Measure-Theoretic Approach to Kernel Conditional Mean Embeddings](#)

- Junhyung Park, Krikamol Muandet
- abstract@[open-review](#): We present a new operator-free, measure-theoretic approach to the conditional mean embedding as a random variable taking values in a reproducing kernel Hilbert space. While the kernel mean embedding of marginal distributions has been defined rigorously, the existing operator-based approach of the conditional version lacks a rigorous treatment, and depends on strong assumptions that hinder its analysis. Our approach does not impose any of the assumptions that the operator-based counterpart requires. We derive a natural regression interpretation to obtain empirical estimates, and provide a thorough analysis of its properties, including universal consistency with improved convergence rates. As natural by-products, we obtain the conditional analogues of the Maximum Mean Discrepancy and Hilbert-Schmidt Independence Criterion, and demonstrate their behaviour via simulations.

[Quantifying the Empirical Wasserstein Distance to a Set of Measures: Beating the Curse of Dimensionality](#)

- Nian Si, Jose Blanchet, Soumyadip Ghosh, Mark Squillante
- abstract@[open-review](#): We consider the problem of estimating the Wasserstein distance between the empirical measure and a set of probability measures whose expectations over a class of functions (hypothesis class) are constrained. If this class is sufficiently rich to characterize a particular distribution (e.g., all Lipschitz functions), then our formulation recovers the Wasserstein distance to such a distribution. We establish a strong duality result that generalizes the celebrated Kantorovich-Rubinstein duality. We also show that our formulation can be used to beat the curse of dimensionality, which is well known to affect the rates of statistical convergence of the empirical Wasserstein distance. In particular, examples of infinite-dimensional hypothesis classes are presented, informed by a complex correlation structure, for which it is shown that the empirical Wasserstein distance to such classes converges to zero at the standard parametric rate. Our formulation provides insights that help clarify why, despite the curse of dimensionality, the Wasserstein distance enjoys favorable empirical performance across a wide range of statistical applications.

[Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning](#)

- Jean-Bastien Grill, Florian Strub, Florent AltchÃ©, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, Michal Valko
- abstract@[open-review](#): We introduce Bootstrap Your Own Latent (BYOL), a new approach to self-supervised image representation learning. BYOL relies on two neural networks, referred to as online and target networks, that interact and learn from each other. From an augmented view of an image, we train the online network to predict the target network representation of the same image under a different augmented view. At the same time, we update the target network with a slow-moving average of the online network. While state-of-the art methods intrinsically rely on negative pairs, BYOL achieves a new state of the art without them. BYOL reaches 74.3% top-1 classification accuracy on ImageNet using the standard linear evaluation protocol with a standard ResNet-50 architecture and 79.6% with a larger ResNet. We also show that BYOL performs on par or better than the current state of the art on both transfer and semi-supervised benchmarks.

[Towards Theoretically Understanding Why Sgd Generalizes Better Than Adam in Deep Learning](#)

- Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, Weinan E

- abstract@[open-review](#): It is not clear yet why ADAM-like adaptive gradient algorithms suffer from worse generalization performance than SGD despite their faster training speed. This work aims to provide understandings on this generalization gap by analyzing their local convergence behaviors. Specifically, we observe the heavy tails of gradient noise in these algorithms. This motivates us to analyze these algorithms through their Levy-driven stochastic differential equations (SDEs) because of the similar convergence behaviors of an algorithm and its SDE. Then we establish the escaping time of these SDEs from a local basin. The result shows that (1) the escaping time of both SGD and ADAM~depends on the Radon measure of the basin positively and the heaviness of gradient noise negatively; (2) for the same basin, SGD enjoys smaller escaping time than ADAM, mainly because (a) the geometry adaptation in ADAM~via adaptively scaling each gradient coordinate well diminishes the anisotropic structure in gradient noise and results in larger Radon measure of a basin; (b) the exponential gradient average in ADAM~smooths its gradient and leads to lighter gradient noise tails than SGD. So SGD is more locally unstable than ADAM~at sharp minima defined as the minima whose local basins have small Radon measure, and can better escape from them to flatter ones with larger Radon measure. As flat minima here which often refer to the minima at flat or asymmetric basins/valleys often generalize better than sharp ones~\cite{keskar2016large,he2019asymmetric}, our result explains the better generalization performance of SGD over ADAM. Finally, experimental results confirm our heavy-tailed gradient noise assumption and theoretical affirmation.

[RSKDD-Net: Random Sample-based Keypoint Detector and Descriptor](#)

- Fan Lu, Guang Chen, Yinlong Liu, Zhongnan Qu, Alois Knoll
- abstract@[open-review](#): Keypoint detector and descriptor are two main components of point cloud registration. Previous learning-based keypoint detectors rely on saliency estimation for each point or farthest point sample (FPS) for candidate points selection, which are inefficient and not applicable in large scale scenes. This paper proposes Random Sample-based Keypoint Detector and Descriptor Network (RSKDD-Net) for large scale point cloud registration. The key idea is using random sampling to efficiently select candidate points and using a learning-based method to jointly generate keypoints and corresponding descriptors. To tackle the information loss of random sampling, we exploit a novel random dilation cluster strategy to enlarge the receptive field of each sampled point and an attention mechanism to aggregate the positions and features of neighbor points. Furthermore, we propose a matching loss to train the descriptor in a weakly supervised manner. Extensive experiments on two large scale outdoor LiDAR datasets show that the proposed RSKDD-Net achieves state-of-the-art performance with more than 15 times faster than existing methods. Our code is available at <https://github.com/ispc-lab/RSKDD-Net>.

[Efficient Clustering for Stretched Mixtures: Landscape and Optimality](#)

- Kaizheng Wang, Yuling Yan, Mateo Diaz
- abstract@[open-review](#): This paper considers a canonical clustering problem where one receives unlabeled samples drawn from a balanced mixture of two elliptical distributions and aims for a classifier to estimate the labels. Many popular methods including PCA and k-means require individual components of the mixture to be somewhat spherical, and perform poorly when they are stretched. To overcome this issue, we propose a non-convex program seeking for an affine transform to turn the data into a one-dimensional point cloud concentrating around -1 and 1, after which clustering becomes easy. Our theoretical contributions are two-fold: (1) we show that the non-convex loss function exhibits desirable geometric properties when the sample size exceeds some constant multiple of the dimension, and (2) we leverage this to prove that an efficient first-order algorithm achieves near-optimal statistical precision without good initialization. We also propose a general methodology for clustering with flexible choices of feature transforms and loss objectives.

[A Group-Theoretic Framework for Data Augmentation](#)

- Shuxiao Chen, Edgar Dobriban, Jane Lee
- abstract@[open-review](#): Data augmentation has become an important part of modern deep learning pipelines and is typically needed to achieve state of the art performance for many learning tasks. It utilizes invariant transformations of the data, such as rotation, scale, and color shift, and the transformed images are added to the training set. However, these transformations are often chosen heuristically and a clear theoretical framework to explain the performance benefits of data augmentation is not available. In this paper, we develop such a framework to explain data augmentation as averaging over the orbits of the group that keeps the data distribution approximately invariant, and show that it leads to variance reduction. We study finite-sample and asymptotic empirical risk minimization and work out as examples the variance reduction in certain two-layer neural networks. We further propose a strategy to exploit the benefits of data augmentation for general learning tasks.

[The Statistical Cost of Robust Kernel Hyperparameter Tuning](#)

- Raphael Meyer, Christopher Musco
- abstract@[open-review](#): This paper studies the statistical complexity of kernel hyperparameter tuning in the setting of active regression under adversarial noise. We consider the problem of finding the best interpolant from a class of kernels with unknown hyperparameters, assuming only that the noise is square-integrable. We provide finite-sample guarantees for the problem, characterizing how increasing the complexity of the kernel class increases the complexity of learning kernel hyperparameters. For common kernel classes (e.g. squared-exponential kernels with unknown lengthscale), our results show that hyperparameter optimization increases sample complexity by just a logarithmic factor, in comparison to the setting where optimal parameters are known in advance. Our result is based on a subsampling guarantee for linear regression under multiple design matrices which may be of independent interest.

[How does Weight Correlation Affect Generalisation Ability of Deep Neural Networks?](#)

- Gaojie Jin, Xinping Yi, Liang Zhang, Lijun Zhang, Sven Schewe, Xiaowei Huang
- abstract@[open-review](#): This paper studies the novel concept of weight correlation in deep neural networks and discusses its impact on the networks' generalisation ability. For fully-connected layers, the weight correlation is defined as the average cosine similarity between weight vectors of neurons, and for convolutional layers, the weight correlation is defined as the cosine similarity between filter matrices. Theoretically, we show that, weight correlation can, and should, be incorporated into the PAC Bayesian framework for the generalisation of neural networks, and the resulting generalisation bound is monotonic with respect to the weight correlation. We formulate a new complexity measure, which lifts the PAC Bayes measure with weight correlation, and experimentally confirm that it is able to rank the generalisation errors of a set of networks more precisely than existing measures. More importantly, we develop a new regulariser for training, and provide extensive experiments that show that the generalisation error can be greatly reduced with our novel approach.

[ContraGAN: Contrastive Learning for Conditional Image Generation](#)

- Minguk Kang, Jaesik Park
- abstract@[open-review](#): Conditional image generation is the task of generating diverse images using class label information. Although many conditional Generative Adversarial Networks (GAN) have shown realistic results, such methods consider pairwise relations between the embedding of an image and the embedding of the corresponding label (data-to-class relations) as the conditioning losses. In this paper, we propose ContraGAN that considers relations between multiple image embeddings in the same batch (data-to-data relations) as well as the data-to-class relations by using a conditional contrastive loss. The discriminator of ContraGAN discriminates the authenticity of given samples and minimizes a contrastive objective to learn the relations between training images. Simultaneously, the generator tries to generate realistic images that deceive the authenticity and have a low contrastive loss. The experimental results show that ContraGAN outperforms state-of-the-art-models by 7.3% and 7.7% on Tiny ImageNet and ImageNet datasets, respectively.

Besides, we experimentally demonstrate that ContraGAN helps to relieve the overfitting of the discriminator. For a fair comparison, we re-implement twelve state-of-the-art GANs using the PyTorch library. The software package is available at <https://github.com/POSTECH-CVLab/PyTorch-StudioGAN>.

[On the distance between two neural networks and the stability of learning](#)

- Jeremy Bernstein, Arash Vahdat, Yisong Yue, Ming-Yu Liu
- abstract@[open-review](#): This paper relates parameter distance to gradient breakdown for a broad class of nonlinear compositional functions. The analysis leads to a new distance function called deep relative trust and a descent lemma for neural networks. Since the resulting learning rule seems to require little to no learning rate tuning, it may unlock a simpler workflow for training deeper and more complex neural networks. The Python code used in this paper is here: <https://github.com/jxbz/fromage>.

[A Topological Filter for Learning with Label Noise](#)

- Pengxiang Wu, Songzhu Zheng, Mayank Goswami, Dimitris Metaxas, Chao Chen
- abstract@[open-review](#): Noisy labels can impair the performance of deep neural networks. To tackle this problem, in this paper, we propose a new method for filtering label noise. Unlike most existing methods relying on the posterior probability of a noisy classifier, we focus on the much richer spatial behavior of data in the latent representational space. By leveraging the high-order topological information of data, we are able to collect most of the clean data and train a high-quality model. Theoretically we prove that this topological approach is guaranteed to collect the clean data with high probability. Empirical results show that our method outperforms the state-of-the-arts and is robust to a broad spectrum of noise types and levels.

[Personalized Federated Learning with Moreau Envelopes](#)

- Canh T. Dinh, Nguyen Tran, Josh Nguyen
- abstract@[open-review](#): Federated learning (FL) is a decentralized and privacy-preserving machine learning technique in which a group of clients collaborate with a server to learn a global model without sharing clients' data. One challenge associated with FL is statistical diversity among clients, which restricts the global model from delivering good performance on each client's task. To address this, we propose an algorithm for personalized FL (pFedMe) using Moreau envelopes as clients' regularized loss functions, which help decouple personalized model optimization from the global model learning in a bi-level problem stylized for personalized FL. Theoretically, we show that pFedMe convergence rate is state-of-the-art: achieving quadratic speedup for strongly convex and sublinear speedup of order 2/3 for smooth nonconvex objectives. Experimentally, we verify that pFedMe excels at empirical performance compared with the vanilla FedAvg and Per-FedAvg, a meta-learning based personalized FL algorithm.

[Avoiding Side Effects in Complex Environments](#)

- Alex Turner, Neale Ratzlaff, Prasad Tadepalli
- abstract@[open-review](#): Reward function specification can be difficult. Rewarding the agent for making a widget may be easy, but penalizing the multitude of possible negative side effects is hard. In toy environments, Attainable Utility Preservation (AUP) avoided side effects by penalizing shifts in the ability to achieve randomly generated goals. We scale this approach to large, randomly generated environments based on Conway's Game of Life. By preserving optimal value for a single randomly generated reward function, AUP incurs modest overhead while leading the agent to complete the specified task and avoid many side effects. Videos and code are available at <https://avoiding-side-effects.github.io/>.

[No-regret Learning in Price Competitions under Consumer Reference Effects](#)

- Negin Golrezaei, Patrick Jaillet, Jason Cheuk Nam Liang
- abstract@[open-review](#): We study long-run market stability for repeated price competitions between two firms, where consumer demand depends on firms' posted prices and consumers' price expectations called reference prices. Consumers' reference prices vary over time according to a memory-based dynamic, which is a weighted average of all historical prices. We focus on the setting where firms are not aware of demand functions and how reference prices are formed but have access to an oracle that provides a measure of consumers' responsiveness to the current posted prices. We show that if the firms run no-regret algorithms, in particular, online mirror descent (OMD), with decreasing step sizes, the market stabilizes in the sense that firms' prices and reference prices converge to a stable Nash Equilibrium (SNE). Interestingly, we also show that there exist constant step sizes under which the market stabilizes. We further characterize the rate of convergence to the SNE for both decreasing and constant OMD step sizes.

[Geometric Dataset Distances via Optimal Transport](#)

- David Alvarez-Melis, Nicolo Fusi
- abstract@[open-review](#): The notion of task similarity is at the core of various machine learning paradigms, such as domain adaptation and meta-learning. Current methods to quantify it are often heuristic, make strong assumptions on the label sets across the tasks, and many are architecture-dependent, relying on task-specific optimal parameters (e.g., require training a model on each dataset). In this work we propose an alternative notion of distance between datasets that (i) is model-agnostic, (ii) does not involve training, (iii) can compare datasets even if their label sets are completely disjoint and (iv) has solid theoretical footing. This distance relies on optimal transport, which provides it with rich geometry awareness, interpretable correspondences and well-understood properties. Our results show that this novel distance provides meaningful comparison of datasets, and correlates well with transfer learning hardness across various experimental settings and datasets.

[Task-Agnostic Amortized Inference of Gaussian Process Hyperparameters](#)

- Sulin Liu, Xingyuan Sun, Peter J. Ramadge, Ryan P. Adams
- abstract@[open-review](#): Gaussian processes (GPs) are flexible priors for modeling functions. However, their success depends on the kernel accurately reflecting the properties of the data. One of the appeals of the GP framework is that the marginal likelihood of the kernel hyperparameters is often available in closed form, enabling optimization and sampling procedures to fit these hyperparameters to data. Unfortunately, point-wise evaluation of the marginal likelihood is expensive due to the need to solve a linear system; searching or sampling the space of hyperparameters thus often dominates the practical cost of using GPs. We introduce an approach to the identification of kernel hyperparameters in GP regression and related problems that sidesteps the need for costly marginal likelihoods. Our strategy is to "amortize" inference over hyperparameters by training a single neural network, which consumes a set of regression data and produces an estimate of the kernel function, useful across different tasks. To accommodate the varying dimension and cardinality of different regression problems, we use a hierarchical self-attention-based neural network that produces estimates of the hyperparameters which are invariant to the order of the input data points and data dimensions. We show that a single neural model trained on synthetic data is able to generalize directly to several different unseen real-world GP use cases. Our experiments demonstrate that the estimated hyperparameters are comparable in quality to those from the conventional model selection procedures, while being much faster to obtain, significantly accelerating GP regression and its related applications such as Bayesian optimization and Bayesian quadrature. The code and pre-trained model are available at <https://github.com/PrincetonLIPS/AHGP>.

[A novel variational form of the Schatten-\\$p\\$ quasi-norm](#)

- Paris Giampouras, Rene Vidal, Athanasios Rontogiannis, Benjamin Haeffele
- abstract@[open-review](#): The Schatten-\$p\$ quasi-norm with \$p \in (0,1)\$ has recently gained considerable attention in various low-rank matrix estimation problems offering significant benefits over relevant convex heuristics such as the nuclear norm. However, due to the nonconvexity of the Schatten-\$p\$ quasi-norm, minimization suffers from two major drawbacks: 1) the lack of theoretical guarantees and 2) the high computational cost which is demanded for the minimization task even for trivial tasks such as finding stationary points. In an attempt to reduce the high computational cost induced by Schatten-\$p\$ quasi-norm minimization, variational forms, which are defined over smaller-size matrix factors whose product equals the original matrix, have been proposed. Here, we propose and analyze a novel {Nit variational form of Schatten-\$p\$ quasi-norm} which, for the first time in the literature, is defined for any continuous value of \$p \in (0,1]\$ and decouples along the columns of the factorized matrices. The proposed form can be considered as the natural generalization of the well-known variational form of the nuclear norm to the nonconvex case i.e., for \$p \in (0,1]\$. Notably, low-rankness is now imposed via a group-sparsity promoting regularizer. The resulting formulation gives way to SVD-free algorithms thus offering lower computational complexity than the one that is induced by the original definition of the Schatten-\$p\$ quasi-norm. A local optimality analysis is provided which shows~that we can arrive at a local minimum of the original Schatten-\$p\$ quasi-norm problem by reaching a local minimum of the matrix factorization based surrogate problem. In addition, for the case of the squared Frobenius loss with linear operators obeying the restricted isometry property (RIP), a rank-one update scheme is proposed, which offers a way to escape poor local minima. Finally, the efficiency of our approach is empirically shown on a matrix completion problem.

[Energy-based Out-of-distribution Detection](#)

- Weitang Liu, Xiaoyun Wang, John Owens, Yixuan Li
- abstract@[open-review](#): Determining whether inputs are out-of-distribution (OOD) is an essential building block for safely deploying machine learning models in the open world. However, previous methods relying on the softmax confidence score suffer from overconfident posterior distributions for OOD data. We propose a unified framework for OOD detection that uses an energy score. We show that energy scores better distinguish in- and out-of-distribution samples than the traditional approach using the softmax scores. Unlike softmax confidence scores, energy scores are theoretically aligned with the probability density of the inputs and are less susceptible to the overconfidence issue. Within this framework, energy can be flexibly used as a scoring function for any pre-trained neural classifier as well as a trainable cost function to shape the energy surface explicitly for OOD detection. On a CIFAR-10 pre-trained WideResNet, using the energy score reduces the average FPR (at TPR 95%) by 18.03% compared to the softmax confidence score. With energy-based training, our method outperforms the state-of-the-art on common benchmarks.

[On the Loss Landscape of Adversarial Training: Identifying Challenges and How to Overcome Them](#)

- Chen Liu, Mathieu Salzmann, Tao Lin, Ryota Tomioka, Sabine Sässstrunk
- abstract@[open-review](#): We analyze the influence of adversarial training on the loss landscape of machine learning models. To this end, we first provide analytical studies of the properties of adversarial loss functions under different adversarial budgets. We then demonstrate that the adversarial loss landscape is less favorable to optimization, due to increased curvature and more scattered gradients. Our conclusions are validated by numerical analyses, which show that training under large adversarial budgets impede the escape from suboptimal random initialization, cause non-vanishing gradients and make the models' minima found sharper. Based on these observations, we show that a periodic adversarial scheduling (PAS) strategy can effectively overcome these challenges, yielding better results than vanilla adversarial training while being much less sensitive to the choice of learning rate.

[User-Dependent Neural Sequence Models for Continuous-Time Event Data](#)

- Alex Boyd, Robert Bamler, Stephan Mandt, Padhraic Smyth
- abstract@[open-review](#): Continuous-time event data are common in applications such as individual behavior data, financial transactions, and medical health records. Modeling such data can be very challenging, in particular for applications with many different types of events, since it requires a model to predict the event types as well as the time of occurrence. Recurrent neural networks that parameterize time-varying intensity functions are the current state-of-the-art for predictive modeling with such data. These models typically assume that all event sequences come from the same data distribution. However, in many applications event sequences are generated by different sources, or users, and their characteristics can be very different. In this paper, we extend the broad class of neural marked point process models to mixtures of latent embeddings, where each mixture component models the characteristic traits of a given user. Our approach relies on augmenting these models with a latent variable that encodes user characteristics, represented by a mixture model over user behavior that is trained via amortized variational inference. We evaluate our methods on four large real-world datasets and demonstrate systematic improvements from our approach over existing work for a variety of predictive metrics such as log-likelihood, next event ranking, and source-of-sequence identification.

[Active Structure Learning of Causal DAGs via Directed Clique Trees](#)

- Chandler Squires, Sara Magliacane, Kristjan Greenewald, Dmitriy Katz, Murat Kocaoglu, Karthikeyan Shanmugam
- abstract@[open-review](#): A growing body of work has begun to study intervention design for efficient structure learning of causal directed acyclic graphs (DAGs). A typical setting is a \text{causally sufficient} setting, i.e. a system with no latent confounders, selection bias, or feedback, when the essential graph of the observational equivalence class (EC) is given as an input and interventions are assumed to be noiseless. Most existing works focus on \text{worst-case} or \text{average-case} lower bounds for the number of interventions required to orient a DAG. These worst-case lower bounds only establish that the largest clique in the essential graph \text{could} make it difficult to learn the true DAG. In this work, we develop a \text{universal} lower bound for single-node interventions that establishes that the largest clique is \text{always} a fundamental impediment to structure learning. Specifically, we present a decomposition of a DAG into independently orientable components through \text{directed clique trees} and use it to prove that the number of single-node interventions necessary to orient any DAG in an EC is at least the sum of half the size of the largest cliques in each chain component of the essential graph. Moreover, we present a two-phase intervention design algorithm that, under certain conditions on the chordal skeleton, matches the optimal number of interventions up to a multiplicative logarithmic factor in the number of maximal cliques. We show via synthetic experiments that our algorithm can scale to much larger graphs than most of the related work and achieves better worst-case performance than other scalable approaches. A code base to recreate these results can be found at \url{https://github.com/csquires/dct-policy}.

[Convergence and Stability of Graph Convolutional Networks on Large Random Graphs](#)

- Nicolas Keriven, Alberto Bietti, Samuel Vaiter
- abstract@[open-review](#): We study properties of Graph Convolutional Networks (GCNs) by analyzing their behavior on standard models of random graphs, where nodes are represented by random latent variables and edges are drawn according to a similarity kernel. This allows us to overcome the difficulties of dealing with discrete notions such as isomorphisms on very large graphs, by considering instead more natural geometric aspects. We first study the convergence of GCNs to their continuous counterpart as the number of nodes grows. Our results are fully non-asymptotic and are valid for relatively sparse graphs with an average degree that grows logarithmically with the number of nodes. We then analyze the stability of GCNs to small deformations of the random graph model. In contrast to previous studies of stability in discrete settings, our continuous setup allows us to provide more intuitive deformation-based metrics for understanding stability, which have proven useful for explaining the success of convolutional representations on Euclidean domains.

[BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization](#)

- Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G. Wilson, Eytan Bakshy
- abstract@[open-review](#): Bayesian optimization provides sample-efficient global optimization for a broad range of applications, including automatic machine learning, engineering, physics, and experimental design. We introduce BoTorch, a modern programming framework for Bayesian optimization that combines Monte-Carlo (MC) acquisition functions, a novel sample average approximation optimization approach, auto-differentiation, and variance reduction techniques. BoTorch's modular design facilitates flexible specification and optimization of probabilistic models written in PyTorch, simplifying implementation of new acquisition functions. Our approach is backed by novel theoretical convergence results and made practical by a distinctive algorithmic foundation that leverages fast predictive distributions, hardware acceleration, and deterministic optimization. We also propose a novel "one-shot" formulation of the Knowledge Gradient, enabled by a combination of our theoretical and software contributions. In experiments, we demonstrate the improved sample efficiency of BoTorch relative to other popular libraries.

[Reconsidering Generative Objectives For Counterfactual Reasoning](#)

- Danni Lu, Chenyang Tao, Junya Chen, Fan Li, Feng Guo, Lawrence Carin
- abstract@[open-review](#): There has been recent interest in exploring generative goals for counterfactual reasoning, such as individualized treatment effect (ITE) estimation. However, existing solutions often fail to address issues that are unique to causal inference, such as covariate balancing and (infeasible) counterfactual validation. As a step towards more flexible, scalable and accurate ITE estimation, we present a novel generative Bayesian estimation framework that integrates representation learning, adversarial matching and causal estimation. By appealing to the Robinson decomposition, we derive a reformulated variational bound that explicitly targets the causal effect estimation rather than specific predictive goals. Our procedure acknowledges the uncertainties in representation and solves a Fenchel mini-max game to resolve the representation imbalance for better counterfactual generalization, justified by new theory. Further, the latent variable formulation employed enables robustness to unobservable latent confounders, extending the scope of its applicability. The utility of the proposed solution is demonstrated via an extensive set of tests against competing solutions, both under various simulation setups and to real-world datasets, with encouraging results reported.

[Robust Federated Learning: The Case of Affine Distribution Shifts](#)

- Amirhossein Reisizadeh, Farzan Farnia, Ramtin Pedarsani, Ali Jadbabaie
- abstract@[open-review](#): Federated learning is a distributed paradigm that aims at training models using samples distributed across multiple users in a network while keeping the samples on users' devices with the aim of efficiency and protecting users privacy. In such settings, the training data is often statistically heterogeneous and manifests various distribution shifts across users, which degrades the performance of the learnt model. The primary goal of this paper is to develop a robust federated learning algorithm that achieves satisfactory performance against distribution shifts in users' samples. To achieve this goal, we first consider a structured affine distribution shift in users' data that captures the device-dependent data heterogeneity in federated settings. This perturbation model is applicable to various federated learning problems such as image classification where the images undergo device-dependent imperfections, e.g. different intensity, contrast, and brightness. To address affine distribution shifts across users, we propose a Federated Learning framework Robust to Affine distribution shifts (FLRA) that is provably robust against affine Wasserstein shifts to the distribution of observed samples. To solve the FLRA's distributed minimax optimization problem, we propose a fast and efficient optimization method and provide convergence and performance guarantees via a gradient Descent Ascent (GDA) method. We further prove generalization error bounds for the learnt classifier to show proper generalization from empirical distribution of samples to the true underlying distribution. We perform several numerical experiments to empirically support FLRA. We show that an affine distribution shift indeed suffices to significantly decrease the performance of the learnt classifier in a new test user, and our proposed algorithm achieves a significant gain in comparison to standard federated learning and adversarial training methods.

[Quantile Propagation for Wasserstein-Approximate Gaussian Processes](#)

- Rui Zhang, Christian Walder, Edwin V. Bonilla, Marian-Andrei Rizoiu, Lexing Xie
- abstract@[open-review](#): Approximate inference techniques are the cornerstone of probabilistic methods based on Gaussian process priors. Despite this, most work approximately optimizes standard divergence measures such as the Kullback-Leibler (KL) divergence, which lack the basic desiderata for the task at hand, while chiefly offering merely technical convenience. We develop a new approximate inference method for Gaussian process models which overcomes the technical challenges arising from abandoning these convenient divergences. Our method---dubbed Quantile Propagation (QP)---is similar to expectation propagation (EP) but minimizes the $\$L_2\$$ Wasserstein distance (WD) instead of the KL divergence. The WD exhibits all the required properties of a distance metric, while respecting the geometry of the underlying sample space. We show that QP matches quantile functions rather than moments as in EP and has the same mean update but a smaller variance update than EP, thereby alleviating EP's tendency to over-estimate posterior variances. Crucially, despite the significant complexity of dealing with the WD, QP has the same favorable locality property as EP, and thereby admits an efficient algorithm. Experiments on classification and Poisson regression show that QP outperforms both EP and variational Bayes.

[Generating Adjacency-Constrained Subgoals in Hierarchical Reinforcement Learning](#)

- Tianren Zhang, Shangqi Guo, Tian Tan, Xiaolin Hu, Feng Chen
- abstract@[open-review](#): Goal-conditioned hierarchical reinforcement learning (HRL) is a promising approach for scaling up reinforcement learning (RL) techniques. However, it often suffers from training inefficiency as the action space of the high-level, i.e., the goal space, is often large. Searching in a large goal space poses difficulties for both high-level subgoal generation and low-level policy learning. In this paper, we show that this problem can be effectively alleviated by restricting the high-level action space from the whole goal space to a k-step adjacent region of the current state using an adjacency constraint. We theoretically prove that the proposed adjacency constraint preserves the optimal hierarchical policy in deterministic MDPs, and show that this constraint can be practically implemented by training an adjacency network that can discriminate between adjacent and non-adjacent subgoals. Experimental results on discrete and continuous control tasks show that incorporating the adjacency constraint improves the performance of state-of-the-art HRL approaches in both deterministic and stochastic environments.

[High-contrast ‘gaudy’ images improve the training of deep neural network models of visual cortex](#)

- Benjamin Cowley, Jonathan W. Pillow
- abstract@[open-review](#): A key challenge in understanding the sensory transformations of the visual system is to obtain a highly predictive model that maps natural images to neural responses. Deep neural networks (DNNs) provide a promising candidate for such a model. However, DNNs require orders of magnitude more training data than neuroscientists can collect because experimental recording time is severely limited. This motivates us to find images that train highly-predictive DNNs with as little training data as possible. We propose high-contrast, binarized versions of natural images---termed gaudy images---to efficiently train DNNs to predict higher-order visual cortical responses. In simulation experiments and analyses of real neural data, we find that training DNNs with gaudy images substantially reduces the number of training images needed to accurately predict responses to natural images. We also find that gaudy images, chosen before training, outperform images chosen during training by active learning algorithms. Thus, gaudy images overemphasize features of natural images that are the most important for efficiently training DNNs. We believe gaudy images will aid in the modeling of visual cortical neurons, potentially opening new scientific questions about visual processing.

[Duality-Induced Regularizer for Tensor Factorization Based Knowledge Graph Completion](#)

- Zhanqiu Zhang, Jianyu Cai, Jie Wang

- abstract@[open-review](#): Tensor factorization based models have shown great power in knowledge graph completion (KGC). However, their performance usually suffers from the overfitting problem seriously. This motivates various regularizers---such as the squared Frobenius norm and tensor nuclear norm regularizers---while the limited applicability significantly limits their practical usage. To address this challenge, we propose a novel regularizer---namely, DURA ---which is not only effective in improving the performance of existing models but widely applicable to various methods. The major novelty of DURA is based on the observation that, for an existing tensor factorization based KGC model (primal), there is often another distance based KGC model (dual) closely associated with it.

[Distributed Training with Heterogeneous Data: Bridging Median- and Mean-Based Algorithms](#)

- Xiangyi Chen, Tiancong Chen, Haoran Sun, Steven Z. Wu, Mingyi Hong
- abstract@[open-review](#): Recently, there is a growing interest in the study of median-based algorithms for distributed non-convex optimization. Two prominent examples include signSGD with majority vote, an effective approach for communication reduction via 1-bit compression on the local gradients, and medianSGD, an algorithm recently proposed to ensure robustness against Byzantine workers. The convergence analyses for these algorithms critically rely on the assumption that all the distributed data are drawn iid from the same distribution. However, in applications such as Federated Learning, the data across different nodes or machines can be inherently heterogeneous, which violates such an iid assumption. This work analyzes signSGD and medianSGD in distributed settings with heterogeneous data. We show that these algorithms are non-convergent whenever there is some disparity between the expected median and mean over the local gradients. To overcome this gap, we provide a novel gradient correction mechanism that perturbs the local gradients with noise, which we show can provably close the gap between mean and median of the gradients. The proposed methods largely preserve nice properties of these median-based algorithms, such as the low per-iteration communication complexity of signSGD, and further enjoy global convergence to stationary solutions. Our perturbation technique can be of independent interest when one wishes to estimate mean through a median estimator.

[H-Mem: Harnessing synaptic plasticity with Hebbian Memory Networks](#)

- Thomas Limbacher, Robert Legenstein
- abstract@[open-review](#): The ability to base current computations on memories from the past is critical for many cognitive tasks such as story understanding. Hebbian-type synaptic plasticity is believed to underlie the retention of memories over medium and long time scales in the brain. However, it is unclear how such plasticity processes are integrated with computations in cortical networks. Here, we propose Hebbian Memory Networks (H-Mems), a simple neural network model that is built around a core hetero-associative network subject to Hebbian plasticity. We show that the network can be optimized to utilize the Hebbian plasticity processes for its computations. H-Mems can one-shot memorize associations between stimulus pairs and use these associations for decisions later on. Furthermore, they can solve demanding question-answering tasks on synthetic stories. Our study shows that neural network models are able to enrich their computations with memories through simple Hebbian plasticity processes.

[Neural Unsigned Distance Fields for Implicit Function Learning](#)

- Julian Chibane, Mohamad Aymen mir, Gerard Pons-Moll
- abstract@[open-review](#): In this work we target a learnable output representation that allows continuous, high resolution outputs of arbitrary shape. Recent works represent 3D surfaces implicitly with a Neural Network, thereby breaking previous barriers in resolution, and ability to represent diverse topologies. However, neural implicit representations are limited to closed surfaces, which divide the space into inside and outside. Many real world objects such as walls of a scene scanned by a sensor, clothing, or a car with inner structures are not closed. This constitutes a significant barrier, in terms of data pre-processing (objects need to be artificially closed creating artifacts), and the ability to output open surfaces. In this work, we propose Neural Distance Fields (NDF), a neural network based model which predicts the unsigned distance field for arbitrary 3D shapes given sparse point clouds. NDF represent surfaces at high resolutions as prior implicit models, but do not require closed surface data, and significantly broaden the class of representable shapes in the output. NDF allow to extract the surface as very dense point clouds and as meshes. We also show that NDF allow for surface normal calculation and can be rendered using a slight modification of sphere tracing. We find NDF can be used for multi-target regression (multiple outputs for one input) with techniques that have been exclusively used for rendering in graphics. Experiments on ShapeNet show that NDF, while simple, is the state-of-the art, and allows to reconstruct shapes with inner structures, such as the chairs inside a bus. Notably, we show that NDF are not restricted to 3D shapes, and can approximate more general open surfaces such as curves, manifolds, and functions. Code is available for research at <https://virtualhumans.mpi-inf.mpg.de/ndf/>.

[Curriculum By Smoothing](#)

- Samarth Sinha, Animesh Garg, Hugo Larochelle
- abstract@[open-review](#): Convolutional Neural Networks (CNNs) have shown impressive performance in computer vision tasks such as image classification, detection, and segmentation. Moreover, recent work in Generative Adversarial Networks (GANs) has highlighted the importance of learning by progressively increasing the difficulty of a learning task Kerras et al. When learning a network from scratch, the information propagated within the network during the earlier stages of training can contain distortion artifacts due to noise which can be detrimental to training. In this paper, we propose an elegant curriculum-based scheme that smoothes the feature embedding of a CNN using anti-aliasing or low-pass filters. We propose to augment the training of CNNs by controlling the amount of high frequency information propagated within the CNNs as training progresses, by convolving the output of a CNN feature map of each layer with a Gaussian kernel. By decreasing the variance of the Gaussian kernel, we gradually increase the amount of high-frequency information available within the network for inference. As the amount of information in the feature maps increases during training, the network is able to progressively learn better representations of the data. Our proposed augmented training scheme significantly improves the performance of CNNs on various vision tasks without either adding additional trainable parameters or an auxiliary regularization objective. The generality of our method is demonstrated through empirical performance gains in CNN architectures across four different tasks: transfer learning, cross-task transfer learning, and generative models.

[Fast Transformers with Clustered Attention](#)

- Apoorv Vyas, Angelos Katharopoulos, FranÃ§ois Fleuret
- abstract@[open-review](#): Transformers have been proven a successful model for a variety of tasks in sequence modeling. However, computing the attention matrix, which is their key component, has quadratic complexity with respect to the sequence length, thus making them prohibitively expensive for large sequences. To address this, we propose clustered attention, which instead of computing the attention for every query, groups queries into clusters and computes attention just for the centroids. To further improve this approximation, we use the computed clusters to identify the keys with the highest attention per query and compute the exact key/query dot products. This results in a model with linear complexity with respect to the sequence length for a fixed number of clusters. We evaluate our approach on two automatic speech recognition datasets and show that our model consistently outperforms vanilla transformers for a given computational budget. Finally, we demonstrate that our model can approximate arbitrarily complex attention distributions with a minimal number of clusters by approximating a pretrained BERT model on GLUE and SQuAD benchmarks with only 25 clusters and no loss in performance.

[The Convex Relaxation Barrier, Revisited: Tightened Single-Neuron Relaxations for Neural Network Verification](#)

- Christian Tjandraatmadja, Ross Anderson, Joey Huchette, Will Ma, KRUNAL KISHOR PATEL, Juan Pablo Vielma

- abstract@[open-review](#): We improve the effectiveness of propagation- and linear-optimization-based neural network verification algorithms with a new tightened convex relaxation for ReLU neurons. Unlike previous single-neuron relaxations which focus only on the univariate input space of the ReLU, our method considers the multivariate input space of the affine pre-activation function preceding the ReLU. Using results from submodularity and convex geometry, we derive an explicit description of the tightest possible convex relaxation when this multivariate input is over a box domain. We show that our convex relaxation is significantly stronger than the commonly used univariate-input relaxation which has been proposed as a natural convex relaxation barrier for verification. While our description of the relaxation may require an exponential number of inequalities, we show that they can be separated in linear time and hence can be efficiently incorporated into optimization algorithms on an as-needed basis. Based on this novel relaxation, we design two polynomial-time algorithms for neural network verification: a linear-programming-based algorithm that leverages the full power of our relaxation, and a fast propagation algorithm that generalizes existing approaches. In both cases, we show that for a modest increase in computational effort, our strengthened relaxation enables us to verify a significantly larger number of instances compared to similar algorithms.

[**Strongly Incremental Constituency Parsing with Graph Neural Networks**](#)

- Kaiyu Yang, Jia Deng
- abstract@[open-review](#): Parsing sentences into syntax trees can benefit downstream applications in NLP. Transition-based parsers build trees by executing actions in a state transition system. They are computationally efficient, and can leverage machine learning to predict actions based on partial trees. However, existing transition-based parsers are predominantly based on the shift-reduce transition system, which does not align with how humans are known to parse sentences. Psycholinguistic research suggests that human parsing is strongly incremental—"humans grow a single parse tree by adding exactly one token at each step. In this paper, we propose a novel transition system called attach-juxtapose. It is strongly incremental; it represents a partial sentence using a single tree; each action adds exactly one token into the partial tree. Based on our transition system, we develop a strongly incremental parser. At each step, it encodes the partial tree using a graph neural network and predicts an action. We evaluate our parser on Penn Treebank (PTB) and Chinese Treebank (CTB). On PTB, it outperforms existing parsers trained with only constituency trees; and it performs on par with state-of-the-art parsers that use dependency trees as additional training data. On CTB, our parser establishes a new state of the art. Code is available at <https://github.com/princeton-vl/attach-juxtapose-parser>.

[**AOT: Appearance Optimal Transport Based Identity Swapping for Forgery Detection**](#)

- Hao Zhu, Chaoyou Fu, Qianyi Wu, Wayne Wu, Chen Qian, Ran He
- abstract@[open-review](#): Recent studies have shown that the performance of forgery detection can be improved with diverse and challenging Deepfakes datasets. However, due to the lack of Deepfakes datasets with large variance in appearance, which can be hardly produced by recent identity swapping methods, the detection algorithm may fail in this situation. In this work, we provide a new identity swapping algorithm with large differences in appearance for face forgery detection. The appearance gaps mainly arise from the large discrepancies in illuminations and skin colors that widely exist in real-world scenarios. However, due to the difficulties of modeling the complex appearance mapping, it is challenging to transfer fine-grained appearances adaptively while preserving identity traits. This paper formulates appearance mapping as an optimal transport problem and proposes an Appearance Optimal Transport model (AOT) to formulate it in both latent and pixel space. Specifically, a relighting generator is designed to simulate the optimal transport plan. It is solved via minimizing the Wasserstein distance of the learned features in the latent space, enabling better performance and less computation than conventional optimization. To further refine the solution of the optimal transport plan, we develop a segmentation game to minimize the Wasserstein distance in the pixel space. A discriminator is introduced to distinguish the fake parts from a mix of real and fake image patches. Extensive experiments reveal that the superiority of our method when compared with state-of-the-art methods and the ability of our generated data to improve the performance of face forgery detection.

[**Uncertainty-Aware Learning for Zero-Shot Semantic Segmentation**](#)

- Ping Hu, Stan Sclaroff, Kate Saenko
- abstract@[open-review](#): Zero-shot semantic segmentation (ZSS) aims to classify pixels of novel classes without training examples available. Recently, most ZSS methods focus on learning the visual-semantic correspondence to transfer knowledge from seen classes to unseen classes at the pixel level. Yet, few works study the adverse effects caused by the noisy and outlying training samples in the seen classes. In this paper, we identify this challenge and address it with a novel framework that learns to discriminate noisy samples based on Bayesian uncertainty estimation. Specifically, we model the network outputs with Gaussian and Laplacian distributions, with the variances accounting for the observation noise as well as the uncertainty of input samples. Learning objectives are then derived with the estimated variances playing as adaptive attenuation for individual samples in training. Consequently, our model learns more attentively from representative samples of seen classes while suffering less from noisy and outlying ones, thus providing better reliability and generalization toward unseen categories. We demonstrate the effectiveness of our framework through comprehensive experiments on multiple challenging benchmarks, and show that our method achieves significant accuracy improvement over previous approaches for large open-set segmentation.

[**Delta-STN: Efficient Bilevel Optimization for Neural Networks using Structured Response Jacobians**](#)

- Juhan Bae, Roger B. Grosse
- abstract@[open-review](#): Hyperparameter optimization of neural networks can be elegantly formulated as a bilevel optimization problem. While research on bilevel optimization of neural networks has been dominated by implicit differentiation and unrolling, hypernetworks such as Self-Tuning Networks (STNs) have recently gained traction due to their ability to amortize the optimization of the inner objective. In this paper, we diagnose several subtle pathologies in the training of STNs. Based on these observations, we propose the Delta-STN, an improved hypernetwork architecture which stabilizes training and optimizes hyperparameters much more efficiently than STNs. The key idea is to focus on accurately approximating the best-response Jacobian rather than the full best-response function; we achieve this by reparameterizing the hypernetwork and linearizing the network around the current parameters. We demonstrate empirically that our Delta-STN can tune regularization hyperparameters (e.g. weight decay, dropout, number of cutout holes) with higher accuracy, faster convergence, and improved stability compared to existing approaches.

[**First-Order Methods for Large-Scale Market Equilibrium Computation**](#)

- Yuan Gao, Christian Kroer
- abstract@[open-review](#): Market equilibrium is a solution concept with many applications such as digital ad markets, fair division, and resource sharing. For many classes of utility functions, equilibria can be captured by convex programs. We develop simple first-order methods suitable for solving these programs for large-scale markets. We focus on three practically-relevant utility classes: linear, quasilinear, and Leontief utilities. Using structural properties of market equilibria under each utility class, we show that the corresponding convex programs can be reformulated as optimization of a structured smooth convex function over a polyhedral set, for which projected gradient achieves linear convergence. To do so, we utilize recent linear convergence results under weakened strong-convexity conditions, and further refine the relevant constants in existing convergence results. Then, we show that proximal gradient (a generalization of projected gradient) with a practical linesearch scheme achieves linear convergence under the Proximal-PL condition, a recently developed error bound condition for convex composite problems. For quasilinear utilities, we show that Mirror Descent applied to a new convex program achieves sublinear last-iterate convergence and yields a form of Proportional Response dynamics, an elegant, interpretable algorithm for computing market equilibria originally developed for linear utilities. Numerical experiments show that Proportional Response is highly efficient for computing approximate market equilibria, while projected gradient with linesearch can be much faster when higher-accuracy solutions are needed.

Minimax Optimal Nonparametric Estimation of Heterogeneous Treatment Effects

- Zijun Gao, Yanjun Han
- abstract@[open-review](#): A central goal of causal inference is to detect and estimate the treatment effects of a given treatment or intervention on an outcome variable of interest, where a member known as the heterogeneous treatment effect (HTE) is of growing popularity in recent practical applications such as the personalized medicine. In this paper, we model the HTE as a smooth nonparametric difference between two less smooth baseline functions, and determine the tight statistical limits of the nonparametric HTE estimation as a function of the covariate geometry. In particular, a two-stage nearest-neighbor-based estimator throwing away observations with poor matching quality is near minimax optimal. We also establish the tight dependence on the density ratio without the usual assumption that the covariate densities are bounded away from zero, where a key step is to employ a novel maximal inequality which could be of independent interest.

Residual Force Control for Agile Human Behavior Imitation and Extended Motion Synthesis

- Ye Yuan, Kris Kitani
- abstract@[open-review](#): Reinforcement learning has shown great promise for synthesizing realistic human behaviors by learning humanoid control policies from motion capture data. However, it is still very challenging to reproduce sophisticated human skills like ballet dance, or to stably imitate long-term human behaviors with complex transitions. The main difficulty lies in the dynamics mismatch between the humanoid model and real humans. That is, motions of real humans may not be physically possible for the humanoid model. To overcome the dynamics mismatch, we propose a novel approach, residual force control (RFC), that augments a humanoid control policy by adding external residual forces into the action space. During training, the RFC-based policy learns to apply residual forces to the humanoid to compensate for the dynamics mismatch and better imitate the reference motion. Experiments on a wide range of dynamic motions demonstrate that our approach outperforms state-of-the-art methods in terms of convergence speed and the quality of learned motions. Notably, we showcase a physics-based virtual character empowered by RFC that can perform highly agile ballet dance moves such as piroette, arabesque and jeté. Furthermore, we propose a dual-policy control framework, where a kinematic policy and an RFC-based policy work in tandem to synthesize multi-modal infinite-horizon human motions without any task guidance or user input. Our approach is the first humanoid control method that successfully learns from a large-scale human motion dataset (Human3.6M) and generates diverse long-term motions. Code and videos are available at <https://www.ye-yuan.com/rfc>.

A General Method for Robust Learning from Batches

- Ayush Jain, Alon Orlitsky
- abstract@[open-review](#): In many applications, data is collected in batches, some of which may be corrupt or even adversarial. Recent work derived optimal robust algorithms for estimating finite distributions in this setting. We develop a general framework of robust learning from batches, and determine the limits of both distribution estimation, and notably, classification, over arbitrary, including continuous, domains. Building on this framework, we derive the first robust agnostic: (1) polynomial-time distribution estimation algorithms for structured distributions, including piecewise-polynomial, monotone, log-concave, and gaussian-mixtures, and also significantly improve their sample complexity; (2) classification algorithms, and also establish their near-optimal sample complexity; (3) computationally-efficient algorithms for the fundamental problem of interval-based classification that underlies nearly all natural-1-dimensional classification problems.

Not All Unlabeled Data are Equal: Learning to Weight Data in Semi-supervised Learning

- Zhongzheng Ren, Raymond Yeh, Alexander Schwing
- abstract@[open-review](#): Existing semi-supervised learning (SSL) algorithms use a single weight to balance the loss of labeled and unlabeled examples, i.e., all unlabeled examples are equally weighted. But not all unlabeled data are equal. In this paper we study how to use a different weight for every unlabeled example. Manual tuning of all those weights -- as done in prior work -- is no longer possible. Instead, we adjust those weights via an algorithm based on the influence function, a measure of a model's dependency on one training example. To make the approach efficient, we propose a fast and effective approximation of the influence function. We demonstrate that this technique outperforms state-of-the-art methods on semi-supervised image and language classification tasks.

Hard Negative Mixing for Contrastive Learning

- Yannis Kalantidis, Mert Bulent Sarıyıldız, Noe Pion, Philippe Weinzaepfel, Diane Larlus
- abstract@[open-review](#): Contrastive learning has become a key component of self-supervised learning approaches for computer vision. By learning to embed two augmented versions of the same image close to each other and to push the embeddings of different images apart, one can train highly transferable visual representations. As revealed by recent studies, heavy data augmentation and large sets of negatives are both crucial in learning such representations. At the same time, data mixing strategies, either at the image or the feature level, improve both supervised and semi-supervised learning by synthesizing novel examples, forcing networks to learn more robust features. In this paper, we argue that an important aspect of contrastive learning, i.e. the effect of hard negatives, has so far been neglected. To get more meaningful negative samples, current top contrastive self-supervised learning approaches either substantially increase the batch sizes, or keep very large memory banks; increasing memory requirements, however, leads to diminishing returns in terms of performance. We therefore start by delving deeper into a top-performing framework and show evidence that harder negatives are needed to facilitate better and faster learning. Based on these observations, and motivated by the success of data mixing, we propose hard negative mixing strategies at the feature level, that can be computed on-the-fly with a minimal computational overhead. We exhaustively ablate our approach on linear classification, object detection, and instance segmentation and show that employing our hard negative mixing procedure improves the quality of visual representations learned by a state-of-the-art self-supervised learning method.

MOReL: Model-Based Offline Reinforcement Learning

- Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, Thorsten Joachims
- abstract@[open-review](#): In offline reinforcement learning (RL), the goal is to learn a highly rewarding policy based solely on a dataset of historical interactions with the environment. This serves as an extreme test for an agent's ability to effectively use historical data which is known to be critical for efficient RL. Prior work in offline RL has been confined almost exclusively to model-free RL approaches. In this work, we present MOReL, an algorithmic framework for model-based offline RL. This framework consists of two steps: (a) learning a pessimistic MDP using the offline dataset; (b) learning a near-optimal policy in this pessimistic MDP. The design of the pessimistic MDP is such that for any policy, the performance in the real environment is approximately lower-bounded by the performance in the pessimistic MDP. This enables the pessimistic MDP to serve as a good surrogate for purposes of policy evaluation and learning. Theoretically, we show that MOReL is minimax optimal (up to log factors) for offline RL. Empirically, MOReL matches or exceeds state-of-the-art results on widely used offline RL benchmarks. Overall, the modular design of MOReL enables translating advances in its components (for e.g., in model learning, planning etc.) to improvements in offline RL.

Weisfeiler and Leman go sparse: Towards scalable higher-order graph embeddings

- Christopher Morris, Gaurav Rattan, Petra Mutzel

- abstract@[open-review](#): Graph kernels based on the $\$1\$$ -dimensional Weisfeiler-Leman algorithm and corresponding neural architectures recently emerged as powerful tools for (supervised) learning with graphs. However, due to the purely local nature of the algorithms, they might miss essential patterns in the given data and can only handle binary relations. The $\$k\$$ -dimensional Weisfeiler-Leman algorithm addresses this by considering $\$k\$$ -tuples, defined over the set of vertices, and defines a suitable notion of adjacency between these vertex tuples. Hence, it accounts for the higher-order interactions between vertices. However, it does not scale and may suffer from overfitting when used in a machine learning setting. Hence, it remains an important open problem to design WL-based graph learning methods that are simultaneously expressive, scalable, and non-overfitting. Here, we propose local variants and corresponding neural architectures, which consider a subset of the original neighborhood, making them more scalable, and less prone to overfitting. The expressive power of (one of) our algorithms is strictly higher than the original algorithm, in terms of ability to distinguish non-isomorphic graphs. Our experimental study confirms that the local algorithms, both kernel and neural architectures, lead to vastly reduced computation times, and prevent overfitting. The kernel version establishes a new state-of-the-art for graph classification on a wide range of benchmark datasets, while the neural version shows promising performance on large-scale molecular regression tasks.

[Adversarial Crowdsourcing Through Robust Rank-One Matrix Completion](#)

- Qianqian Ma, Alex Olshevsky
- abstract@[open-review](#): We consider the problem of reconstructing a rank-one matrix from a revealed subset of its entries when some of the revealed entries are corrupted with perturbations that are unknown and can be arbitrarily large. It is not known which revealed entries are corrupted. We propose a new algorithm combining alternating minimization with extreme-value filtering and provide sufficient and necessary conditions to recover the original rank-one matrix. In particular, we show that our proposed algorithm is optimal when the set of revealed entries is given by an Erdős-Rényi random graph. These results are then applied to the problem of classification from crowdsourced data under the assumption that while the majority of the workers are governed by the standard single-coin David-Skene model (i.e., they output the correct answer with a certain probability), some of the workers can deviate arbitrarily from this model. In particular, the ``adversarial'' workers could even make decisions designed to make the algorithm output an incorrect answer. Extensive experimental results show our algorithm for this problem, based on rank-one matrix completion with perturbations, outperforms all other state-of-the-art methods in such an adversarial scenario.

[Learning Semantic-aware Normalization for Generative Adversarial Networks](#)

- Heliang Zheng, Jianlong Fu, Yanhong Zeng, Jiebo Luo, Zheng-Jun Zha
- abstract@[open-review](#): The recent advances in image generation have been achieved by style-based image generators. Such approaches learn to disentangle latent factors in different image scales and encode latent factors as style to control image synthesis. However, existing approaches cannot further disentangle fine-grained semantics from each other, which are often conveyed from feature channels. In this paper, we propose a novel image synthesis approach by learning Semantic-aware relative importance for feature channels in Generative Adversarial Networks (SariGAN). Such a model disentangles latent factors according to the semantic of feature channels by channel-/group- wise fusion of latent codes and feature channels. Particularly, we learn to cluster feature channels by semantics and propose an adaptive group-wise Normalization (AdaGN) to independently control the styles of different channel groups. For example, we can adjust the statistics of channel groups for a human face to control the open and close of the mouth, while keeping other facial features unchanged. We propose to use adversarial training, a channel grouping loss, and a mutual information loss for joint optimization, which not only enables high-fidelity image synthesis but leads to superior interpretable properties. Extensive experiments show that our approach outperforms the SOTA style-based approaches in both unconditional image generation and conditional image inpainting tasks.

[Differentiable Causal Discovery from Interventional Data](#)

- Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, Alexandre Drouin
- abstract@[open-review](#): Learning a causal directed acyclic graph from data is a challenging task that involves solving a combinatorial problem for which the solution is not always identifiable. A new line of work reformulates this problem as a continuous constrained optimization one, which is solved via the augmented Lagrangian method. However, most methods based on this idea do not make use of interventional data, which can significantly alleviate identifiability issues. This work constitutes a new step in this direction by proposing a theoretically-grounded method based on neural networks that can leverage interventional data. We illustrate the flexibility of the continuous-constrained framework by taking advantage of expressive neural architectures such as normalizing flows. We show that our approach compares favorably to the state of the art in a variety of settings, including perfect and imperfect interventions for which the targeted nodes may even be unknown.

[One-sample Guided Object Representation Disassembling](#)

- Zunlei Feng, Yongming He, Xinchao Wang, Xin Gao, Jie Lei, Cheng Jin, Mingli Song
- abstract@[open-review](#): The ability to disassemble the features of objects and background is crucial for many machine learning tasks, including image classification, image editing, visual concepts learning, and so on. However, existing (semi-)supervised methods all need a large amount of annotated samples, while unsupervised methods can't handle real-world images with complicated backgrounds. In this paper, we introduce the One-sample Guided Object Representation Disassembling (One-GORD) method, which only requires one annotated sample for each object category to learn disassembled object representation from unannotated images. For the annotated one-sample, we first adopt some data augmentation strategies to generate some synthetic samples, which can guide the disassembling of the object features and background features. For the unannotated images, two self-supervised mechanisms: dual-swapping and fuzzy classification are introduced to disassemble object features from the background with the guidance of annotated one-sample. What's more, we devise two metrics to evaluate the disassembling performance from the perspective of representation and image, respectively. Experiments demonstrate that the One-GORD achieves competitive disassembling performance and can handle natural scenes with complicated backgrounds.

[Extrapolation Towards Imaginary 0-Nearest Neighbour and Its Improved Convergence Rate](#)

- Akifumi Okuno, Hidetoshi Shimodaira
- abstract@[open-review](#): $\$k\$$ -nearest neighbour ($\$k\$$ -NN) is one of the simplest and most widely-used methods for supervised classification, that predicts a query's label by taking weighted ratio of observed labels of $\$k\$$ objects nearest to the query. The weights and the parameter $\$k \in \mathbb{N}$ regulate its bias-variance trade-off, and the trade-off implicitly affects the convergence rate of the excess risk for the $\$k\$$ -NN classifier; several existing studies considered selecting optimal $\$k\$$ and weights to obtain faster convergence rate. Whereas $\$k\$$ -NN with non-negative weights has been developed widely, it was also proved that negative weights are essential for eradicating the bias terms and attaining optimal convergence rate. In this paper, we propose a novel multiscale $\$k\$$ -NN (MS- $\$k\$$ -NN), that extrapolates unweighted $\$k\$$ -NN estimators from several $\$k \geq 1\$$ values to $\$k=0\$$, thus giving an imaginary 0-NN estimator. Our method implicitly computes optimal real-valued weights that are adaptive to the query and its neighbour points. We theoretically prove that the MS- $\$k\$$ -NN attains the improved rate, which coincides with the existing optimal rate under some conditions.

[Robust Persistence Diagrams using Reproducing Kernels](#)

- Siddharth Vishwanath, Kenji Fukumizu, Satoshi Kuriki, Bharath K. Sriperumbudur
- abstract@[open-review](#): Persistent homology has become an important tool for extracting geometric and topological features from data, whose multi-scale features are summarized in a persistence diagram. From a statistical perspective, however, persistence diagrams are very sensitive to perturbations in the input space. In this work, we develop a framework for constructing robust persistence diagrams from superlevel filtrations of robust density estimators

constructed using reproducing kernels. Using an analogue of the influence function on the space of persistence diagrams, we establish the proposed framework to be less sensitive to outliers. The robust persistence diagrams are shown to be consistent estimators in the bottleneck distance, with the convergence rate controlled by the smoothness of the kernel — this, in turn, allows us to construct uniform confidence bands in the space of persistence diagrams. Finally, we demonstrate the superiority of the proposed approach on benchmark datasets.

[Contextual Games: Multi-Agent Learning with Side Information](#)

- Pier Giuseppe Sessa, Ilija Bogunovic, Andreas Krause, Maryam Kamgarpour
- abstract@[open-review](#): We formulate the novel class of contextual games, a type of repeated games driven by contextual information at each round. By means of kernel-based regularity assumptions, we model the correlation between different contexts and game outcomes and propose a novel online (meta) algorithm that exploits such correlations to minimize the contextual regret of individual players. We define game-theoretic notions of contextual Coarse Correlated Equilibria (c-CCE) and optimal contextual welfare for this new class of games and show that c-CCEs and optimal welfare can be approached whenever players' contextual regrets vanish. Finally, we empirically validate our results in a traffic routing experiment, where our algorithm leads to better performance and higher welfare compared to baselines that do not exploit the available contextual information or the correlations present in the game.

[Goal-directed Generation of Discrete Structures with Conditional Generative Models](#)

- Amina Mollaysa, Brooks Paige, Alexandros Kalousis
- abstract@[open-review](#): Despite recent advances, goal-directed generation of structured discrete data remains challenging. For problems such as program synthesis (generating source code) and materials design (generating molecules), finding examples which satisfy desired constraints or exhibit desired properties is difficult. In practice, expensive heuristic search or reinforcement learning algorithms are often employed. In this paper, we investigate the use of conditional generative models which directly attack this inverse problem, by modeling the distribution of discrete structures given properties of interest. Unfortunately, the maximum likelihood training of such models often fails with the samples from the generative model inadequately respecting the input properties. To address this, we introduce a novel approach to directly optimize a reinforcement learning objective, maximizing an expected reward. We avoid high-variance score-function estimators that would otherwise be required by sampling from an approximation to the normalized rewards, allowing simple Monte Carlo estimation of model gradients. We test our methodology on two tasks: generating molecules with user-defined properties and identifying short python expressions which evaluate to a given target value. In both cases, we find improvements over maximum likelihood estimation and other baselines.

[Beyond Lazy Training for Over-parameterized Tensor Decomposition](#)

- Xiang Wang, Chenwei Wu, Jason D. Lee, Tengyu Ma, Rong Ge
- abstract@[open-review](#): Over-parametrization is an important technique in training neural networks. In both theory and practice, training a larger network allows the optimization algorithm to avoid bad local optimal solutions. In this paper we study a closely related tensor decomposition problem: given an $\$l$ -th order tensor in $(\mathbb{R}^d)^{\otimes l}$ of rank r (where $r \ll d$), can variants of gradient descent find a rank m decomposition where $m > r$? We show that in a lazy training regime (similar to the NTK regime for neural networks) one needs at least $m = \Omega(d^{1-\epsilon})$, while a variant of gradient descent can find an approximate tensor when $m = O^*(r^{2.5}\log d)$. Our results show that gradient descent on over-parametrized objective could go beyond the lazy training regime and utilize certain low-rank structure in the data.

[Denoised Smoothing: A Provable Defense for Pretrained Classifiers](#)

- Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, J. Zico Kolter
- abstract@[open-review](#): We present a method for provably defending any pretrained image classifier against ℓ_p adversarial attacks. This method, for instance, allows public vision API providers and users to seamlessly convert pretrained non-robust classification services into provably robust ones. By prepending a custom-trained denoiser to any off-the-shelf image classifier and using randomized smoothing, we effectively create a new classifier that is guaranteed to be ℓ_p -robust to adversarial examples, without modifying the pretrained classifier. Our approach applies to both the white-box and the black-box settings of the pretrained classifier. We refer to this defense as denoised smoothing, and we demonstrate its effectiveness through extensive experimentation on ImageNet and CIFAR-10. Finally, we use our approach to provably defend the Azure, Google, AWS, and ClarifAI image classification APIs. Our code replicating all the experiments in the paper can be found at: <https://github.com/microsoft/denoised-smoothing>.

[Minibatch Stochastic Approximate Proximal Point Methods](#)

- Hilal Asi, Karan Chadha, Gary Cheng, John C. Duchi
- abstract@[open-review](#): We extend the Approximate-Proximal Point (aProx) family of model-based methods for solving stochastic convex optimization problems, including stochastic subgradient, proximal point, and bundle methods, to the minibatch setting. To do this, we propose two minibatched algorithms for which we prove a non-asymptotic upper bound on the rate of convergence, revealing a linear speedup in minibatch size. In contrast to standard stochastic gradient methods, these methods may have linear speedup in the minibatch setting even for non-smooth functions. Our algorithms maintain the desirable traits characteristic of the aProx family, such as robustness to initial step size choice. Additionally, we show improved convergence rates for "interpolation" problems, which (for example) gives a new parallelization strategy for alternating projections. We corroborate our theoretical results with extensive empirical testing, which demonstrates the gains provided by accurate modeling and minibatching.

[Attribute Prototype Network for Zero-Shot Learning](#)

- Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, Zeynep Akata
- abstract@[open-review](#): From the beginning of zero-shot learning research, visual attributes have been shown to play an important role. In order to better transfer attribute-based knowledge from known to unknown classes, we argue that an image representation with integrated attribute localization ability would be beneficial for zero-shot learning. To this end, we propose a novel zero-shot representation learning framework that jointly learns discriminative global and local features using only class-level attributes. While a visual-semantic embedding layer learns global features, local features are learned through an attribute prototype network that simultaneously regresses and decorrelates attributes from intermediate features. We show that our locality augmented image representations achieve a new state-of-the-art on three zero-shot learning benchmarks. As an additional benefit, our model points to the visual evidence of the attributes in an image, e.g. for the CUB dataset, confirming the improved attribute localization ability of our image representation.

[CrossTransformers: spatially-aware few-shot transfer](#)

- Carl Doersch, Ankush Gupta, Andrew Zisserman
- abstract@[open-review](#): Given new tasks with very little data—such as new classes in a classification problem or a domain shift in the input—performance of modern vision systems degrades remarkably quickly. In this work, we illustrate how the neural network representations which underpin modern vision systems are subject to supervision collapse, whereby they lose any information that is not necessary for performing the training task, including information that may be necessary for transfer to new tasks or domains. We then propose two methods to mitigate this problem. First, we employ self-supervised learning to encourage general-purpose features that transfer better. Second, we propose a novel Transformer based neural network architecture called CrossTransformers, which can take a small number of labeled images and an unlabeled query, find coarse spatial correspondence between the query

and the labeled images, and then infer class membership by computing distances between spatially-corresponding features. The result is a classifier that is more robust to task and domain shift, which we demonstrate via state-of-the-art performance on Meta-Dataset, a recent dataset for evaluating transfer from ImageNet to many other vision datasets.

[Learning Latent Space Energy-Based Prior Model](#)

- Bo Pang, Tian Han, Erik Nijkamp, Song-Chun Zhu, Ying Nian Wu
- abstract@[open-review](#): We propose an energy-based model (EBM) in the latent space of a generator model, so that the EBM serves as a prior model that stands on the top-down network of the generator model. Both the latent space EBM and the top-down network can be learned jointly by maximum likelihood, which involves short-run MCMC sampling from both the prior and posterior distributions of the latent vector. Due to the low dimensionality of the latent space and the expressiveness of the top-down network, a simple EBM in latent space can capture regularities in the data effectively, and MCMC sampling in latent space is efficient and mixes well. We show that the learned model exhibits strong performances in terms of image and text generation and anomaly detection. The one-page code can be found in supplementary materials.

[SEVIR : A Storm Event Imagery Dataset for Deep Learning Applications in Radar and Satellite Meteorology](#)

- Mark Veillette, Siddharth Samsi, Chris Mattioli
- abstract@[open-review](#): Modern deep learning approaches have shown promising results in meteorological applications like precipitation nowcasting, synthetic radar generation, front detection and several others. In order to effectively train and validate these complex algorithms, large and diverse datasets containing high-resolution imagery are required. Petabytes of weather data, such as from the Geostationary Environmental Satellite System (GOES) and the Next-Generation Radar (NEXRAD) system, are available to the public; however, the size and complexity of these datasets is a hindrance to developing and training deep models. To help address this problem, we introduce the Storm EVent ImagRy (SEVIR) dataset - a single, rich dataset that combines spatially and temporally aligned data from multiple sensors, along with baseline implementations of deep learning models and evaluation metrics, to accelerate new algorithmic innovations. SEVIR is an annotated, curated and spatio-temporally aligned dataset containing over 10,000 weather events that each consist of 384 km x 384 km image sequences spanning 4 hours of time. Images in SEVIR were sampled and aligned across five different data types: three channels (C02, C09, C13) from the GOES-16 advanced baseline imager, NEXRAD vertically integrated liquid mosaics, and GOES-16 Geostationary Lightning Mapper (GLM) flashes. Many events in SEVIR were selected and matched to the NOAA Storm Events database so that additional descriptive information such as storm impacts and storm descriptions can be linked to the rich imagery provided by the sensors. We describe the data collection methodology and illustrate the applications of this dataset with two examples of deep learning in meteorology: precipitation nowcasting and synthetic weather radar generation. In addition, we also describe a set of metrics that can be used to evaluate the outputs of these models. The SEVIR dataset and baseline implementations of selected applications are available for download.

[Lightweight Generative Adversarial Networks for Text-Guided Image Manipulation](#)

- Bowen Li, Xiaojuan Qi, Philip Torr, Thomas Lukasiewicz
- abstract@[open-review](#): We propose a novel lightweight generative adversarial network for efficient image manipulation using natural language descriptions. To achieve this, a new word-level discriminator is proposed, which provides the generator with fine-grained training feedback at word-level, to facilitate training a lightweight generator that has a small number of parameters, but can still correctly focus on specific visual attributes of an image, and then edit them without affecting other contents that are not described in the text. Furthermore, thanks to the explicit training signal related to each word, the discriminator can also be simplified to have a lightweight structure. Compared with the state of the art, our method has a much smaller number of parameters, but still achieves a competitive manipulation performance. Extensive experimental results demonstrate that our method can better disentangle different visual attributes, then correctly map them to corresponding semantic words, and thus achieve a more accurate image modification using natural language descriptions.

[High-Dimensional Contextual Policy Search with Unknown Context Rewards using Bayesian Optimization](#)

- Qing Feng , Ben Letham, Hongzi Mao, Eytan Bakshy
- abstract@[open-review](#): Contextual policies are used in many settings to customize system parameters and actions to the specifics of a particular setting. In some real-world settings, such as randomized controlled trials or A/B tests, it may not be possible to measure policy outcomes at the level of contextâ€”we observe only aggregate rewards across a distribution of contexts. This makes policy optimization much more difficult because we must solve a high-dimensional optimization problem over the entire space of contextual policies, for which existing optimization methods are not suitable. We develop effective models that leverage the structure of the search space to enable contextual policy optimization directly from the aggregate rewards using Bayesian optimization. We use a collection of simulation studies to characterize the performance and robustness of the models, and show that our approach of inferring a low-dimensional context embedding performs best. Finally, we show successful contextual policy optimization in a real-world video bitrate policy problem.

[Model Fusion via Optimal Transport](#)

- Sidak Pal Singh, Martin Jaggi
- abstract@[open-review](#): Combining different models is a widely used paradigm in machine learning applications. While the most common approach is to form an ensemble of models and average their individual predictions, this approach is often rendered infeasible by given resource constraints in terms of memory and computation, which grow linearly with the number of models. We present a layer-wise model fusion algorithm for neural networks that utilizes optimal transport to (soft-) align neurons across the models before averaging their associated parameters. We show that this can successfully yield "one-shot" knowledge transfer (i.e., without requiring any retraining) between neural networks trained on heterogeneous non-i.i.d. data. In both i.i.d. and non-i.i.d. settings, we illustrate that our approach significantly outperforms vanilla averaging, as well as how it can serve as an efficient replacement for the ensemble with moderate fine-tuning, for standard convolutional networks (like VGG11), residual networks (like ResNet18), and multi-layer perceptrons on CIFAR10, CIFAR100, and MNIST. Finally, our approach also provides a principled way to combine the parameters of neural networks with different widths, and we explore its application for model compression. The code is available at the following link, <https://github.com/sidak/otfusion>.

[On the Stability and Convergence of Robust Adversarial Reinforcement Learning: A Case Study on Linear Quadratic Systems](#)

- Kaiqing Zhang, Bin Hu, Tamer Basar
- abstract@[open-review](#): Reinforcement learning (RL) algorithms can fail to generalize due to the gap between the simulation and the real world. One standard remedy is to use robust adversarial RL (RARL) that accounts for this gap during the policy training, by modeling the gap as an adversary against the training agent. In this work, we reexamine the effectiveness of RARL under a fundamental robust control setting: the linear quadratic (LQ) case. We first observe that the popular RARL scheme that greedily alternates agentsâ€™ updates can easily destabilize the system. Motivated by this, we propose several other policy-based RARL algorithms whose convergence behaviors are then studied both empirically and theoretically. We find: i) the conventional RARL framework (Pinto et al., 2017) can learn a destabilizing policy if the initial policy does not enjoy the robust stability property against the adversary; and ii) with robustly stabilizing initializations, our proposed double-loop RARL algorithm provably converges to the global optimal cost while maintaining robust stability on-the-fly. We also examine the stability and convergence issues of other variants of policy-based RARL algorithms, and then discuss several ways to learn robustly stabilizing initializations. From a robust control perspective, we aim to provide some new and critical

angles about RARL, by identifying and addressing the stability issues in this fundamental LQ setting in continuous control. Our results make an initial attempt toward better theoretical understandings of policy-based RARL, the core approach in Pinto et al., 2017.

[Learning Individually Inferred Communication for Multi-Agent Cooperation](#)

- Ziluo Ding, Tiejun Huang, Zongqing Lu
- abstract@[open-review](#): Communication lays the foundation for human cooperation. It is also crucial for multi-agent cooperation. However, existing work focuses on broadcast communication, which is not only impractical but also leads to information redundancy that could even impair the learning process. To tackle these difficulties, we propose Individually Inferred Communication (I2C), a simple yet effective model to enable agents to learn a prior for agent-agent communication. The prior knowledge is learned via causal inference and realized by a feed-forward neural network that maps the agent's local observation to a belief about who to communicate with. The influence of one agent on another is inferred via the joint action-value function in multi-agent reinforcement learning and quantified to label the necessity of agent-agent communication. Furthermore, the agent policy is regularized to better exploit communicated messages. Empirically, we show that I2C can not only reduce communication overhead but also improve the performance in a variety of multi-agent cooperative scenarios, comparing to existing methods.

[Set2Graph: Learning Graphs From Sets](#)

- Hadar Serviansky, Nimrod Segol, Jonathan Shlomi, Kyle Cranmer, Eilam Gross, Haggai Maron, Yaron Lipman
- abstract@[open-review](#): Many problems in machine learning (ML) can be cast as learning functions from sets to graphs, or more generally to hypergraphs; in short, Set2Graph functions. Examples include clustering, learning vertex and edge features on graphs, and learning features on triplets in a collection. A natural approach for building Set2Graph models is to characterize all linear equivariant set-to-hypergraph layers and stack them with non-linear activations. This poses two challenges: (i) the expressive power of these networks is not well understood; and (ii) these models would suffer from high, often intractable computational and memory complexity, as their dimension grows exponentially. This paper advocates a family of neural network models for learning Set2Graph functions that is both practical and of maximal expressive power (universal), that is, can approximate arbitrary continuous Set2Graph functions over compact sets. Testing these models on different machine learning tasks, mainly an application to particle physics, we find them favorable to existing baselines.

[Graph Random Neural Networks for Semi-Supervised Learning on Graphs](#)

- Wenzheng Feng, Jie Zhang, Yuxiao Dong, Yu Han, Huanbo Luan, Qian Xu, Qiang Yang, Evgeny Kharlamov, Jie Tang
- abstract@[open-review](#): We study the problem of semi-supervised learning on graphs, for which graph neural networks (GNNs) have been extensively explored. However, most existing GNNs inherently suffer from the limitations of over-smoothing, non-robustness, and weak-generalization when labeled nodes are scarce. In this paper, we propose a simple yet effective framework—"GRAPH RANDOM NEURAL NETWORKS (GRAND)" to address these issues. In GRAND, we first design a random propagation strategy to perform graph data augmentation. Then we leverage consistency regularization to optimize the prediction consistency of unlabeled nodes across different data augmentations. Extensive experiments on graph benchmark datasets suggest that GRAND significantly outperforms state-of-the-art GNN baselines on semi-supervised node classification. Finally, we show that GRAND mitigates the issues of over-smoothing and non-robustness, exhibiting better generalization behavior than existing GNNs. The source code of GRAND is publicly available at <https://github.com/Grand20/grand>.

[Gradient Boosted Normalizing Flows](#)

- Robert Giaquinto, Arindam Banerjee
- abstract@[open-review](#): By chaining a sequence of differentiable invertible transformations, normalizing flows (NF) provide an expressive method of posterior approximation, exact density evaluation, and sampling. The trend in normalizing flow literature has been to devise deeper, more complex transformations to achieve greater flexibility. We propose an alternative: Gradient Boosted Normalizing Flows (GBNF) model a density by successively adding new NF components with gradient boosting. Under the boosting framework, each new NF component optimizes a weighted likelihood objective, resulting in new components that are fit to the suitable residuals of the previously trained components. The GBNF formulation results in a mixture model structure, whose flexibility increases as more components are added. Moreover, GBNFs offer a wider, as opposed to strictly deeper, approach that improves existing NFs at the cost of additional training---not more complex transformations. We demonstrate the effectiveness of this technique for density estimation and, by coupling GBNF with a variational autoencoder, generative modeling of images. Our results show that GBNFs outperform their non-boosted analog, and, in some cases, produce better results with smaller, simpler flows.

[Open Graph Benchmark: Datasets for Machine Learning on Graphs](#)

- Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, Jure Leskovec
- abstract@[open-review](#): We present the Open Graph Benchmark (OGB), a diverse set of challenging and realistic benchmark datasets to facilitate scalable, robust, and reproducible graph machine learning (ML) research. OGB datasets are large-scale, encompass multiple important graph ML tasks, and cover a diverse range of domains, ranging from social and information networks to biological networks, molecular graphs, source code ASTs, and knowledge graphs. For each dataset, we provide a unified evaluation protocol using meaningful application-specific data splits and evaluation metrics. In addition to building the datasets, we also perform extensive benchmark experiments for each dataset. Our experiments suggest that OGB datasets present significant challenges of scalability to large-scale graphs and out-of-distribution generalization under realistic data splits, indicating fruitful opportunities for future research. Finally, OGB provides an automated end-to-end graph ML pipeline that simplifies and standardizes the process of graph data loading, experimental setup, and model evaluation. OGB will be regularly updated and welcomes inputs from the community. OGB datasets as well as data loaders, evaluation scripts, baseline code, and leaderboards are publicly available at <https://ogb.stanford.edu> .

[Towards Understanding Hierarchical Learning: Benefits of Neural Representations](#)

- Minshuo Chen, Yu Bai, Jason D. Lee, Tuo Zhao, Huan Wang, Caiming Xiong, Richard Socher
- abstract@[open-review](#): Deep neural networks can empirically perform efficient hierarchical learning, in which the layers learn useful representations of the data. However, how they make use of the intermediate representations are not explained by recent theories that relate them to ``shallow learners'' such as kernels. In this work, we demonstrate that intermediate \emph{neural representations} add more flexibility to neural networks and can be advantageous over raw inputs. We consider a fixed, randomly initialized neural network as a representation function fed into another trainable network. When the trainable network is the quadratic Taylor model of a wide two-layer network, we show that neural representation can achieve improved sample complexities compared with the raw input: For learning a low-rank degree-\$p\$ polynomial (\$p \geq 4\$) in \$d\$ dimension, neural representation requires only \$\widetilde{O}(d^{\lceil p/2 \rceil})\$ samples, while the best-known sample complexity upper bound for the raw input is \$\widetilde{O}(d^{p-1})\$. We contrast our result with a lower bound showing that neural representations do not improve over the raw input (in the infinite width limit), when the trainable network is instead a neural tangent kernel. Our results characterize when neural representations are beneficial, and may provide a new perspective on why depth is important in deep learning.

[Texture Interpolation for Probing Visual Perception](#)

- Jonathan Vacher, Aida Davila, Adam Kohn, Ruben Coen-Cagli
- abstract@[open-review](#): Texture synthesis models are important tools for understanding visual processing. In particular, statistical approaches based on neurally relevant features have been instrumental in understanding aspects of visual perception and of neural coding. New deep learning-based approaches further improve the quality of synthetic textures. Yet, it is still unclear why deep texture synthesis performs so well, and applications of this new framework to probe visual perception are scarce. Here, we show that distributions of deep convolutional neural network (CNN) activations of a texture are well described by elliptical distributions and therefore, following optimal transport theory, constraining their mean and covariance is sufficient to generate new texture samples. Then, we propose the natural geodesics (ie the shortest path between two points) arising with the optimal transport metric to interpolate between arbitrary textures. Compared to other CNN-based approaches, our interpolation method appears to match more closely the geometry of texture perception, and our mathematical framework is better suited to study its statistical nature. We apply our method by measuring the perceptual scale associated to the interpolation parameter in human observers, and the neural sensitivity of different areas of visual cortex in macaque monkeys.

[Hierarchical Neural Architecture Search for Deep Stereo Matching](#)

- Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, Zongyuan Ge
- abstract@[open-review](#): To reduce the human efforts in neural network design, Neural Architecture Search (NAS) has been applied with remarkable success to various high-level vision tasks such as classification and semantic segmentation. The underlying idea for the NAS algorithm is straightforward, namely, to allow the network the ability to choose among a set of operations (eg convolution with different filter sizes), one is able to find an optimal architecture that is better adapted to the problem at hand. However, so far the success of NAS has not been enjoyed by low-level geometric vision tasks such as stereo matching. This is partly due to the fact that state-of-the-art deep stereo matching networks, designed by humans, are already sheer in size. Directly applying the NAS to such massive structures is computationally prohibitive based on the currently available mainstream computing resources. In this paper, we propose the first \emph{end-to-end} hierarchical NAS framework for deep stereo matching by incorporating task-specific human knowledge into the neural architecture search framework. Specifically, following the gold standard pipeline for deep stereo matching (ie, feature extraction -- feature volume construction and dense matching), we optimize the architectures of the entire pipeline jointly. Extensive experiments show that our searched network outperforms all state-of-the-art deep stereo matching architectures and is ranked at the top 1 accuracy on KITTI stereo 2012, 2015, and Middlebury benchmarks, as well as the top 1 on SceneFlow dataset with a substantial improvement on the size of the network and the speed of inference. Code available at <https://github.com/XuelianCheng/LEAStereo>.

[MuSCLE: Multi Sweep Compression of LiDAR using Deep Entropy Models](#)

- Sourav Biswas, Jerry Liu, Kelvin Wong, Shenlong Wang, Raquel Urtasun
- abstract@[open-review](#): We present a novel compression algorithm for reducing the storage of LiDAR sensory data streams. Our model exploits spatio-temporal relationships across multiple LiDAR sweeps to reduce the bitrate of both geometry and intensity values. Towards this goal, we propose a novel conditional entropy model that models the probabilities of the octree symbols, by considering both coarse level geometry and previous sweepsâ™ geometric and intensity information. We then exploit the learned probability to encode the full data-stream into a compact one. Our experiments demonstrate that our method significantly reduces the joint geometry and intensity bitrate over prior state-of-the-art LiDAR compression methods, with a reduction of 7â€“17% and 15â€“35% on the UrbanCity and SemanticKITTI datasets respectively.

[Implicit Bias in Deep Linear Classification: Initialization Scale vs Training Accuracy](#)

- Edward Moroshko, Blake E. Woodworth, Suriya Gunasekar, Jason D. Lee, Nati Srebro, Daniel Soudry
- abstract@[open-review](#): We provide a detailed asymptotic study of gradient flow trajectories and their implicit optimization bias when minimizing the exponential loss over "diagonal linear networks". This is the simplest model displaying a transition between "kernel" and non-kernel ("rich" or "active") regimes. We show how the transition is controlled by the relationship between the initialization scale and how accurately we minimize the training loss. Our results indicate that some limit behavior of gradient descent only kick in at ridiculous training accuracies (well beyond 10^{-100}). Moreover, the implicit bias at reasonable initialization scales and training accuracies is more complex and not captured by these limits.

[Focus of Attention Improves Information Transfer in Visual Features](#)

- Matteo Tiezzi, Stefano Melacci, Alessandro Betti, Marco Maggini, Marco Gori
- abstract@[open-review](#): Unsupervised learning from continuous visual streams is a challenging problem that cannot be naturally and efficiently managed in the classic batch-mode setting of computation. The information stream must be carefully processed accordingly to an appropriate spatio-temporal distribution of the visual data, while most approaches of learning commonly assume uniform probability density. In this paper we focus on unsupervised learning for transferring visual information in a truly online setting by using a computational model that is inspired to the principle of least action in physics. The maximization of the mutual information is carried out by a temporal process which yields online estimation of the entropy terms. The model, which is based on second-order differential equations, maximizes the information transfer from the input to a discrete space of symbols related to the visual features of the input, whose computation is supported by hidden neurons. In order to better structure the input probability distribution, we use a human-like focus of attention model that, coherently with the information maximization model, is also based on second-order differential equations. We provide experimental results to support the theory by showing that the spatio-temporal filtering induced by the focus of attention allows the system to globally transfer more information from the input stream over the focused areas and, in some contexts, over the whole frames with respect to the unfiltered case that yields uniform probability distributions.

[Auditing Differentially Private Machine Learning: How Private is Private SGD?](#)

- Matthew Jagielski, Jonathan Ullman, Alina Oprea
- abstract@[open-review](#): We investigate whether Differentially Private SGD offers better privacy in practice than what is guaranteed by its state-of-the-art analysis. We do so via novel data poisoning attacks, which we show correspond to realistic privacy attacks. While previous work (Ma et al., arXiv 2019) proposed this connection between differential privacy and data poisoning as a defense against data poisoning, our use as a tool for understanding the privacy of a specific mechanism is new. More generally, our work takes a quantitative, empirical approach to understanding the privacy afforded by specific implementations of differentially private algorithms that we believe has the potential to complement and influence analytical work on differential privacy.

[A Dynamical Central Limit Theorem for Shallow Neural Networks](#)

- Zhengdao Chen, Grant Rotskoff, Joan Bruna, Eric Vanden-Eijnden
- abstract@[open-review](#): Recent theoretical work has characterized the dynamics and convergence properties for wide shallow neural networks trained via gradient descent; the asymptotic regime in which the number of parameters tends towards infinity has been dubbed the "mean-field" limit. At initialization, the randomly sampled parameters lead to a deviation from the mean-field limit that is dictated by the classical central limit theorem (CLT). However, the dynamics of training introduces correlations among the parameters raising the question of how the fluctuations evolve during training. Here, we analyze the mean-field dynamics as a Wasserstein gradient flow and prove that the deviations from the mean-field evolution scaled by the width, in the width-asymptotic limit, remain bounded throughout training. This observation has implications for both the approximation rate and the generalization: the upper bound we obtain is controlled by a Monte-Carlo type resampling error, which importantly does not depend on dimension. We also relate the bound on the fluctuations to the total variation norm of the measure to which the dynamics converges, which in turn controls the generalization error.

Measuring Systematic Generalization in Neural Proof Generation with Transformers

- Nicolas Gontier, Koustuv Sinha, Siva Reddy, Chris Pal
- abstract@[open-review](#): We are interested in understanding how well Transformer language models (TLMs) can perform reasoning tasks when trained on knowledge encoded in the form of natural language. We investigate their systematic generalization abilities on a logical reasoning task in natural language, which involves reasoning over relationships between entities grounded in first-order logical proofs. Specifically, we perform soft theorem-proving by leveraging TLMs to generate natural language proofs. We test the generated proofs for logical consistency, along with the accuracy of the final inference. We observe length-generalization issues when evaluated on longer-than-trained sequences. However, we observe TLMs improve their generalization performance after being exposed to longer, exhaustive proofs. In addition, we discover that TLMs are able to generalize better using backward-chaining proofs compared to their forward-chaining counterparts, while they find it easier to generate forward chaining proofs. We observe that models that are not trained to generate proofs are better at generalizing to problems based on longer proofs. This suggests that Transformers have efficient internal reasoning strategies that are harder to interpret. These results highlight the systematic generalization behavior of TLMs in the context of logical reasoning, and we believe this work motivates deeper inspection of their underlying reasoning strategies.

Big Self-Supervised Models are Strong Semi-Supervised Learners

- Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, Geoffrey E. Hinton
- abstract@[open-review](#): One paradigm for learning from few labeled examples while making best use of a large amount of unlabeled data is unsupervised pretraining followed by supervised fine-tuning. Although this paradigm uses unlabeled data in a task-agnostic way, in contrast to common approaches to semi-supervised learning for computer vision, we show that it is surprisingly effective for semi-supervised learning on ImageNet. A key ingredient of our approach is the use of big (deep and wide) networks during pretraining and fine-tuning. We find that, the fewer the labels, the more this approach (task-agnostic use of unlabeled data) benefits from a bigger network. After fine-tuning, the big network can be further improved and distilled into a much smaller one with little loss in classification accuracy by using the unlabeled examples for a second time, but in a task-specific way. The proposed semi-supervised learning algorithm can be summarized in three steps: unsupervised pretraining of a big ResNet model using SimCLRv2, supervised fine-tuning on a few labeled examples, and distillation with unlabeled examples for refining and transferring the task-specific knowledge. This procedure achieves 73.9% ImageNet top-1 accuracy with just 1% of the labels ($\$ \leq \13 labeled images per class) using ResNet-50, a 10X improvement in label efficiency over the previous state-of-the-art. With 10% of labels, ResNet-50 trained with our method achieves 77.5% top-1 accuracy, outperforming standard supervised training with all of the labels.

Learning from Label Proportions: A Mutual Contamination Framework

- Clayton Scott, Jianxin Zhang
- abstract@[open-review](#): Learning from label proportions (LLP) is a weakly supervised setting for classification in which unlabeled training instances are grouped into bags, and each bag is annotated with the proportion of each class occurring in that bag. Prior work on LLP has yet to establish a consistent learning procedure, nor does there exist a theoretically justified, general purpose training criterion. In this work we address these two issues by posing LLP in terms of mutual contamination models (MCMs), which have recently been applied successfully to study various other weak supervision settings. In the process, we establish several novel technical results for MCMs, including unbiased losses and generalization error bounds under non-iid sampling plans. We also point out the limitations of a common experimental setting for LLP, and propose a new one based on our MCM framework.

Fast Matrix Square Roots with Applications to Gaussian Processes and Bayesian Optimization

- Geoff Pleiss, Martin Jankowiak, David Eriksson, Anil Damle, Jacob Gardner
- abstract@[open-review](#): Matrix square roots and their inverses arise frequently in machine learning, e.g., when sampling from high-dimensional Gaussians $N(0, K)$ or $\text{whitening } b$ against covariance matrix K . While existing methods typically require $O(N^3)$ computation, we introduce a highly-efficient quadratic-time algorithm for computing $K^{1/2}b$, $K^{-1/2}b$, and their derivatives through matrix-vector multiplication (MVMs). Our method combines Krylov subspace methods with a rational approximation and typically achieves 4 decimal places of accuracy with fewer than 100 MVMs. Moreover, the backward pass requires little additional computation. We demonstrate our method's applicability on matrices as large as 50,000 by 50,000 - well beyond traditional methods - with little approximation error. Applying this increased scalability to variational Gaussian processes, Bayesian optimization, and Gibbs sampling results in more powerful models with higher accuracy. In particular, we perform variational GP inference with up to 10,000 inducing points and perform Gibbs sampling on a 25,000-dimensional problem.

Self-Adaptively Learning to Demoiré from Focused and Defocused Image Pairs

- Lin Liu, Shanxin Yuan, Jianzhuang Liu, Liping Bao, Gregory Slabaugh, Qi Tian
- abstract@[open-review](#): Moiré artifacts are common in digital photography, resulting from the interference between high-frequency scene content and the color filter array of the camera. Existing deep learning-based demoiré methods trained on large scale datasets are limited in handling various complex moiré patterns, and mainly focus on demoiré of photos taken of digital displays. Moreover, obtaining moiré-free ground-truth in natural scenes is difficult but needed for training. In this paper, we propose a self-adaptive learning method for demoiré of a high-frequency image, with the help of an additional defocused moiré-free blur image. Given an image degraded with moiré artifacts and a moiré-free blur image, our network predicts a moiré-free clean image and a blur kernel with a self-adaptive strategy that does not require an explicit training stage, instead performing test-time adaptation. Our model has two sub-networks and works iteratively. During each iteration, one sub-network takes the moiré image as input, removing moiré patterns and restoring image details, and the other sub-network estimates the blur kernel from the blur image. The two sub-networks are jointly optimized. Extensive experiments demonstrate that our method outperforms state-of-the-art methods and can produce high-quality demoiré results. It can generalize well to the task of removing moiré artifacts caused by display screens. In addition, we build a new moiré dataset, including images with screen and texture moiré artifacts. As far as we know, this is the first dataset with real texture moiré patterns.

Confounding-Robust Policy Evaluation in Infinite-Horizon Reinforcement Learning

- Nathan Kallus, Angela Zhou
- abstract@[open-review](#): Off-policy evaluation of sequential decision policies from observational data is necessary in applications of batch reinforcement learning such as education and healthcare. In such settings, however, unobserved variables confound observed actions, rendering exact evaluation of new policies impossible, i.e, unidentifiable. We develop a robust approach that estimates sharp bounds on the (unidentifiable) value of a given policy in an infinite-horizon problem given data from another policy with unobserved confounding, subject to a sensitivity model. We consider stationary unobserved confounding and compute bounds by optimizing over the set of all stationary state-occupancy ratios that agree with a new partially identified estimating equation and the sensitivity model. We prove convergence to the sharp bounds as we collect more confounded data. Although checking set membership is a linear program, the support function is given by a difficult nonconvex optimization problem. We develop approximations based on nonconvex projected gradient descent and demonstrate the resulting bounds empirically.

Model Class Reliance for Random Forests

- Gavin Smith, Roberto Mansilla, James Goulding

- abstract@[open-review](#): Variable Importance (VI) has traditionally been cast as the process of estimating each variable's contribution to a predictive model's overall performance. Analysis of a single model instance, however, guarantees no insight into a variable's relevance to underlying generative processes. Recent research has sought to address this concern via analysis of Rashomon sets - sets of alternative model instances that exhibit equivalent predictive performance to some reference model, but which take different functional forms. Measures such as Model Class Reliance (MCR) have been proposed, that are computed against Rashomon sets, in order to ascertain how much a variable must be relied on to make robust predictions, or whether alternatives exist. If MCR range is tight, we have no choice but to use a variable; if range is high then there exists competing, perhaps fairer models, that provide alternative explanations of the phenomena being examined. Applications are wide, from enabling construction of 'fairer' models in areas such as recidivism, health analytics and ethical marketing. Tractable estimation of MCR for non-linear models is currently restricted to Kernel Regression under squared loss \cite{fisher2019all}. In this paper we introduce a new technique that extends computation of Model Class Reliance (MCR) to Random Forest classifiers and regressors. The proposed approach addresses a number of open research questions, and in contrast to prior Kernel SVM MCR estimation, runs in linearithmic rather than polynomial time. Taking a fundamentally different approach to previous work, we provide a solution for this important model class, identifying situations where irrelevant covariates do not improve predictions.

[Follow the Perturbed Leader: Optimism and Fast Parallel Algorithms for Smooth Minimax Games](#)

- Arun Suggala, Praneeth Netrapalli
- abstract@[open-review](#): We consider the problem of online learning and its application to solving minimax games. For the online learning problem, Follow the Perturbed Leader (FTPL) is a widely studied algorithm which enjoys the optimal $\mathcal{O}(T^{1/2})$ worst-case regret guarantee for both convex and nonconvex losses. In this work, we show that when the sequence of loss functions is predictable, a simple modification of FTPL which incorporates optimism can achieve better regret guarantees, while retaining the optimal worst-case regret guarantee for unpredictable sequences. A key challenge in obtaining these tighter regret bounds is the stochasticity and optimism in the algorithm, which requires different analysis techniques than those commonly used in the analysis of FTPL. The key ingredient we utilize in our analysis is the dual view of perturbation as regularization. While our algorithm has several applications, we consider the specific application of minimax games. For solving smooth convex-concave games, our algorithm only requires access to a linear optimization oracle. For Lipschitz and smooth nonconvex-nonconcave games, our algorithm requires access to an optimization oracle which computes the perturbed best response. In both these settings, our algorithm solves the game up to an accuracy of $\mathcal{O}(T^{-1/2})$ using T calls to the optimization oracle. An important feature of our algorithm is that it is highly parallelizable and requires only $\mathcal{O}(T^{1/2})$ iterations, with each iteration making $\mathcal{O}(T^{1/2})$ parallel calls to the optimization oracle.

[Agnostic \$\mathcal{Q}\$ -learning with Function Approximation in Deterministic Systems: Near-Optimal Bounds on Approximation Error and Sample Complexity](#)

- Simon S. Du, Jason D. Lee, Gaurav Mahajan, Ruosong Wang
- abstract@[open-review](#): The current paper studies the problem of agnostic \mathcal{Q} -learning with function approximation in deterministic systems where the optimal \mathcal{Q} -function is approximable by a function in the class \mathcal{F} with approximation error $\delta \geq 0$. We propose a novel recursion-based algorithm and show that if $\delta = O(\rho \sqrt{\dim_E})$, then one can find the optimal policy using $O(\dim_E)$ trajectories, where ρ is the gap between the optimal \mathcal{Q} -value of the best actions and that of the second-best actions and \dim_E is the Eluder dimension of \mathcal{F} . Our result has two implications: In conjunction with the lower bound in [Du et al., 2020], our upper bound suggests that the condition $\delta = \tilde{\Theta}(\rho \sqrt{\dim_E})$ is necessary and sufficient for algorithms with polynomial sample complexity. In conjunction with the obvious lower bound in the tabular case, our upper bound suggests that the sample complexity $\tilde{\Theta}(\dim_E)$ is tight in the agnostic setting. Therefore, we help address the open problem on agnostic \mathcal{Q} -learning proposed in [Wen and Van Roy, 2013]. We further extend our algorithm to the stochastic reward setting and obtain similar results.

[Learning to Adapt to Evolving Domains](#)

- Hong Liu, Mingsheng Long, Jianmin Wang, Yu Wang
- abstract@[open-review](#): Domain adaptation aims at knowledge transfer from a labeled source domain to an unlabeled target domain. Current domain adaptation methods have made substantial advances in adapting discrete domains. However, this can be unrealistic in real-world applications, where target data usually comes in an online and continually evolving manner as small batches, posing challenges to classic domain adaptation paradigm: (1) Mainstream domain adaptation methods are tailored to stationary target domains, and can fail in non-stationary environments. (2) Since the target data arrive online, the agent should also maintain competence on previous target domains, i.e. to adapt without forgetting. To tackle these challenges, we propose a meta-adaptation framework which enables the learner to adapt to continually evolving target domain without catastrophic forgetting. Our framework comprises of two components: a meta-objective of learning representations to adapt to evolving domains, enabling meta-learning for unsupervised domain adaptation; and a meta-adapter for learning to adapt without forgetting, reserving knowledge from previous target data. Experiments validate the effectiveness of our method on evolving target domains.

[Synthesizing Tasks for Block-based Programming](#)

- Umair Ahmed, Maria Christakis, Aleksandr Efremov, Nigel Fernandez, Ahana Ghosh, Abhik Roychoudhury, Adish Singla
- abstract@[open-review](#): Block-based visual programming environments play a critical role in introducing computing concepts to K-12 students. One of the key pedagogical challenges in these environments is in designing new practice tasks for a student that match a desired level of difficulty and exercise specific programming concepts. In this paper, we formalize the problem of synthesizing visual programming tasks. In particular, given a reference visual task task^{in} and its solution code code^{in} , we propose a novel methodology to automatically generate a set $\{(\text{task}^{\text{out}}, \text{code}^{\text{out}})\}$ of new tasks along with solution codes such that tasks task^{in} and task^{out} are conceptually similar but visually dissimilar. Our methodology is based on the realization that the mapping from the space of visual tasks to their solution codes is highly discontinuous; hence, directly mutating reference task task^{in} to generate new tasks is futile. Our task synthesis algorithm operates by first mutating code code^{in} to obtain a set of codes $\{\text{code}^{\text{out}}\}$. Then, the algorithm performs symbolic execution over a code code^{out} to obtain a visual task task^{out} ; this step uses the Monte Carlo Tree Search (MCTS) procedure to guide the search in the symbolic tree. We demonstrate the effectiveness of our algorithm through an extensive empirical evaluation and user study on reference tasks taken from the Hour of Code: Classic Maze challenge by Code.org and the Intro to Programming with Karel course by CodeHS.com.

[Scalable Belief Propagation via Relaxed Scheduling](#)

- Vitalii Aksenov, Dan Alistarh, Janne H. Korhonen
- abstract@[open-review](#): The ability to leverage large-scale hardware parallelism has been one of the key enablers of the accelerated recent progress in machine learning. Consequently, there has been considerable effort invested into developing efficient parallel variants of classic machine learning algorithms. However, despite the wealth of knowledge on parallelization, some classic machine learning algorithms often prove hard to parallelize efficiently while maintaining convergence.

In this paper, we focus on efficient parallel algorithms for the key machine learning task of inference on graphical models, in particular on the fundamental belief propagation algorithm. We address the challenge of efficiently parallelizing this classic paradigm by showing how to leverage scalable relaxed schedulers, which reduce parallelization overheads, in this context. We investigate the overheads of relaxation analytically, and present an extensive empirical study, showing that

our approach outperforms previous parallel belief propagation implementations both in terms of scalability and in terms of wall-clock convergence time, on a range of practical applications.

[Firefly Neural Architecture Descent: a General Approach for Growing Neural Networks](#)

- Lemeng Wu, Bo Liu, Peter Stone, Qiang Liu
- abstract@[open-review](#): We propose firefly neural architecture descent, a general framework for progressively and dynamically growing neural networks to jointly optimize the networks' parameters and architectures. Our method works in a steepest descent fashion, which iteratively finds the best network within a functional neighborhood of the original network that includes a diverse set of candidate network structures. By using Taylor approximation, the optimal network structure in the neighborhood can be found with a greedy selection procedure. We show that firefly descent can flexibly grow networks both wider and deeper, and can be applied to learn accurate but resource-efficient neural architectures that avoid catastrophic forgetting in continual learning. Empirically, firefly descent achieves promising results on both neural architecture search and continual learning. In particular, on a challenging continual image classification task, it learns networks that are smaller in size but have higher average accuracy than those learned by the state-of-the-art methods.

[Risk-Sensitive Reinforcement Learning: Near-Optimal Risk-Sample Tradeoff in Regret](#)

- Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, Qiaomin Xie
- abstract@[open-review](#): We study risk-sensitive reinforcement learning in episodic Markov decision processes with unknown transition kernels, where the goal is to optimize the total reward under the risk measure of exponential utility. We propose two provably efficient model-free algorithms, Risk-Sensitive Value Iteration (RSVI) and Risk-Sensitive Q-learning (RSQ). These algorithms implement a form of risk-sensitive optimism in the face of uncertainty, which adapts to both risk-seeking and risk-averse modes of exploration. We prove that RSVI attains an $\mathcal{O}(\lambda \beta^2 H^2 \sqrt{S^3 T})$ regret, while RSQ attains an $\mathcal{O}(\lambda \beta^2 \sqrt{H^4 S T})$ regret, where $\lambda(u) = (e^{3u}-1)/u$ for $u > 0$. In the above, β is the risk parameter of the exponential utility function, S the number of states, A the number of actions, T the total number of timesteps, and H the episode length. On the flip side, we establish a regret lower bound showing that the exponential dependence on β and H is unavoidable for any algorithm with an $\tilde{\mathcal{O}}(T)$ regret (even when the risk objective is on the same scale as the original reward), thus certifying the near-optimality of the proposed algorithms. Our results demonstrate that incorporating risk awareness into reinforcement learning necessitates an exponential cost in β and H , which quantifies the fundamental tradeoff between risk sensitivity (related to aleatoric uncertainty) and sample efficiency (related to epistemic uncertainty). To the best of our knowledge, this is the first regret analysis of risk-sensitive reinforcement learning with the exponential utility.

[Learning to Decode: Reinforcement Learning for Decoding of Sparse Graph-Based Channel Codes](#)

- Salman Habib, Allison Beemer, Joerg Kliwer
- abstract@[open-review](#): We show in this work that reinforcement learning can be successfully applied to decoding short to moderate length sparse graph-based channel codes. Specifically, we focus on low-density parity check (LDPC) codes, which for example have been standardized in the context of 5G cellular communication systems due to their excellent error correcting performance. These codes are typically decoded via belief propagation iterative decoding on the corresponding bipartite (Tanner) graph of the code via flooding, i.e., all check and variable nodes in the Tanner graph are updated at once. In contrast, in this paper we utilize a sequential update policy which selects the optimum check node (CN) scheduling in order to improve decoding performance. In particular, we model the CN update process as a multi-armed bandit process with dependent arms and employ a Q-learning scheme for optimizing the CN scheduling policy. In order to reduce the learning complexity, we propose a novel graph-induced CN clustering approach to partition the state space in such a way that dependencies between clusters are minimized. Our results show that compared to other decoding approaches from the literature, the proposed reinforcement learning scheme not only significantly improves the decoding performance, but also reduces the decoding complexity dramatically once the scheduling policy is learned.

[Faster DBSCAN via subsampled similarity queries](#)

- Heinrich Jiang, Jennifer Jang, Jakub Lacki
- abstract@[open-review](#): DBSCAN is a popular density-based clustering algorithm. It computes the ϵ -neighborhood graph of a dataset and uses the connected components of the high-degree nodes to decide the clusters. However, the full neighborhood graph may be too costly to compute with a worst-case complexity of $O(n^2)$. In this paper, we propose a simple variant called SNG-DBSCAN, which clusters based on a subsampled ϵ -neighborhood graph, only requires access to similarity queries for pairs of points and in particular avoids any complex data structures which need the embeddings of the data points themselves. The runtime of the procedure is $O(sn^2)$, where s is the sampling rate. We show under some natural theoretical assumptions that $s \approx \log n/n$ is sufficient for statistical cluster recovery guarantees leading to an $O(n \log n)$ complexity. We provide an extensive experimental analysis showing that on large datasets, one can subsample as little as 0.1% of the neighborhood graph, leading to as much as over 200x speedup and 250x reduction in RAM consumption compared to scikit-learn's implementation of DBSCAN, while still maintaining competitive clustering performance.

[De-Anonymizing Text by Fingerprinting Language Generation](#)

- Zhen Sun, Roei Schuster, Vitaly Shmatikov
- abstract@[open-review](#): Components of machine learning systems are not (yet) perceived as security hotspots. Secure coding practices, such as ensuring that no execution paths depend on confidential inputs, have not yet been adopted by ML developers. We initiate the study of code security of ML systems by investigating how nucleus sampling---a popular approach for generating text, used for applications such as auto-completion---unwittingly leaks texts typed by users. Our main result is that the series of nucleus sizes for many natural English word sequences is a unique fingerprint. We then show how an attacker can infer typed text by measuring these fingerprints via a suitable side channel (e.g., cache access times), explain how this attack could help de-anonymize anonymous texts, and discuss defenses.

[Multiparameter Persistence Image for Topological Machine Learning](#)

- Mathieu Carrière, Andrew Blumberg
- abstract@[open-review](#): In the last decade, there has been increasing interest in topological data analysis, a new methodology for using geometric structures in data for inference and learning. A central theme in the area is the idea of persistence, which in its most basic form studies how measures of shape change as a scale parameter varies. There are now a number of frameworks that support statistics and machine learning in this context. However, in many applications there are several different parameters one might wish to vary: for example, scale and density. In contrast to the one-parameter setting, techniques for applying statistics and machine learning in the setting of multiparameter persistence are not well understood due to the lack of a concise representation of the results. We introduce a new descriptor for multiparameter persistence, which we call the Multiparameter Persistence Image, that is suitable for machine learning and statistical frameworks, is robust to perturbations in the data, has finer resolution than existing descriptors based on slicing, and can be efficiently computed on data sets of realistic size. Moreover, we demonstrate its efficacy by comparing its performance to other multiparameter descriptors on several classification tasks.

PLANS: Neuro-Symbolic Program Learning from Videos

- Raphaël Dang-Nhu
- abstract@[open-review](#): Recent years have seen the rise of statistical program learning based on neural models as an alternative to traditional rule-based systems for programming by example. Rule-based approaches offer correctness guarantees in an unsupervised way as they inherently capture logical rules, while neural models are more realistically scalable to raw, high-dimensional input, and provide resistance to noisy I/O specifications. We introduce PLANS (Program LeArning from Neurally inferred Specifications), a hybrid model for program synthesis from visual observations that gets the best of both worlds, relying on (i) a neural architecture trained to extract abstract, high-level information from each raw individual input (ii) a rule-based system using the extracted information as I/O specifications to synthesize a program capturing the different observations. In order to address the key challenge of making PLANS resistant to noise in the network's output, we introduce a dynamic filtering algorithm for I/O specifications based on selective classification techniques. We obtain state-of-the-art performance at program synthesis from diverse demonstration videos in the Karel and ViZDoom environments, while requiring no ground-truth program for training.

Matrix Inference and Estimation in Multi-Layer Models

- Parthe Pandit, Mojtaba Sahraee Ardakan, Sundeep Rangan, Philip Schniter, Alyson K. Fletcher
- abstract@[open-review](#): We consider the problem of estimating the input and hidden variables of a stochastic multi-layer neural network from an observation of the output. The hidden variables in each layer are represented as matrices with statistical interactions along both rows as well as columns. This problem applies to matrix imputation, signal recovery via deep generative prior models, multi-task and mixed regression, and learning certain classes of two-layer neural networks. We extend a recently-developed algorithm -- Multi-Layer Vector Approximate Message Passing (ML-VAMP), for this matrix-valued inference problem. It is shown that the performance of the proposed Multi-Layer Matrix VAMP (ML-Mat-VAMP) algorithm can be exactly predicted in a certain random large-system limit, where the dimensions $N \times d$ of the unknown quantities grow as $N \rightarrow \infty$ with d fixed. In the two-layer neural-network learning problem, this scaling corresponds to the case where the number of input features, as well as training samples, grow to infinity but the number of hidden nodes stays fixed. The analysis enables a precise prediction of the parameter and test error of the learning.

MeshSDF: Differentiable Iso-Surface Extraction

- Edoardo Remelli, Artem Lukoianov, Stephan Richter, Benoit Guillard, Timur Bagautdinov, Pierre Baque, Pascal Fua
- abstract@[open-review](#): Geometric Deep Learning has recently made striking progress with the advent of continuous Deep Implicit Fields. They allow for detailed modeling of watertight surfaces of arbitrary topology while not relying on a 3D Euclidean grid, resulting in a learnable parameterization that is not limited in resolution. Unfortunately, these methods are often not suitable for applications that require an explicit mesh-based surface representation because converting an implicit field to such a representation relies on the Marching Cubes algorithm, which cannot be differentiated with respect to the underlying implicit field. In this work, we remove this limitation and introduce a differentiable way to produce explicit surface mesh representations from Deep Signed Distance Functions. Our key insight is that by reasoning on how implicit field perturbations impact local surface geometry, one can ultimately differentiate the 3D location of surface samples with respect to the underlying deep implicit field. We exploit this to define MeshSDF, an end-to-end differentiable mesh representation which can vary its topology. We use two different applications to validate our theoretical insight: Single-View Reconstruction via Differentiable Rendering and Physically-Driven Shape Optimization. In both cases our differentiable parameterization gives us an edge over state-of-the-art algorithms.

Variational Interaction Information Maximization for Cross-domain Disentanglement

- HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, Kee-Eung Kim
- abstract@[open-review](#): Cross-domain disentanglement is the problem of learning representations partitioned into domain-invariant and domain-specific representations, which is a key to successful domain transfer or measuring semantic distance between two domains. Grounded in information theory, we cast the simultaneous learning of domain-invariant and domain-specific representations as a joint objective of multiple information constraints, which does not require adversarial training or gradient reversal layers. We derive a tractable bound of the objective and propose a generative model named Interaction Information Auto-Encoder (IIAE). Our approach reveals insights on the desirable representation for cross-domain disentanglement and its connection to Variational Auto-Encoder (VAE). We demonstrate the validity of our model in the image-to-image translation and the cross-domain retrieval tasks. We further show that our model achieves the state-of-the-art performance in the zero-shot sketch based image retrieval task, even without external knowledge.

Provably Efficient Exploration for Reinforcement Learning Using Unsupervised Learning

- Fei Feng, Ruosong Wang, Wotao Yin, Simon S. Du, Lin Yang
- abstract@[open-review](#): Motivated by the prevailing paradigm of using unsupervised learning for efficient exploration in reinforcement learning (RL) problems [tang2017exploration,bellemare2016unifying], we investigate when this paradigm is provably efficient. We study episodic Markov decision processes with rich observations generated from a small number of latent states. We present a general algorithmic framework that is built upon two components: an unsupervised learning algorithm and a no-regret tabular RL algorithm. Theoretically, we prove that as long as the unsupervised learning algorithm enjoys a polynomial sample complexity guarantee, we can find a near-optimal policy with sample complexity polynomial in the number of latent states, which is significantly smaller than the number of observations. Empirically, we instantiate our framework on a class of hard exploration problems to demonstrate the practicality of our theory.

Faithful Embeddings for Knowledge Base Queries

- Haitian Sun, Andrew Arnold, Tania Bedrax Weiss, Fernando Pereira, William W. Cohen
- abstract@[open-review](#): The deductive closure of an ideal knowledge base (KB) contains exactly the logical queries that the KB can answer. However, in practice KBs are both incomplete and over-specified, failing to answer some queries that have real-world answers. \emph{Query embedding} (QE) techniques have been recently proposed where KB entities and KB queries are represented jointly in an embedding space, supporting relaxation and generalization in KB inference. However, experiments in this paper show that QE systems may disagree with deductive reasoning on answers that do not require generalization or relaxation. We address this problem with a novel QE method that is more faithful to deductive reasoning, and show that this leads to better performance on complex queries to incomplete KBs. Finally we show that inserting this new QE module into a neural question-answering system leads to substantial improvements over the state-of-the-art.

Wasserstein Distances for Stereo Disparity Estimation

- Divyansh Garg, Yan Wang, Bharath Hariharan, Mark Campbell, Kilian Q. Weinberger, Wei-Lun Chao
- abstract@[open-review](#): Existing approaches to depth or disparity estimation output a distribution over a set of pre-defined discrete values. This leads to inaccurate results when the true depth or disparity does not match any of these values. The fact that this distribution is usually learned indirectly through a regression loss causes further problems in ambiguous regions around object boundaries. We address these issues using a new neural network architecture that is capable of outputting arbitrary depth values, and a new loss function that is derived from the Wasserstein distance between the true and the predicted distributions. We validate our approach on a variety of tasks, including stereo disparity and depth estimation, and the downstream 3D object

detection. Our approach drastically reduces the error in ambiguous regions, especially around object boundaries that greatly affect the localization of objects in 3D, achieving the state-of-the-art in 3D object detection for autonomous driving.

[Multi-agent Trajectory Prediction with Fuzzy Query Attention](#)

- Nitin Kamra, Hao Zhu, Dweep Kumarbhai Trivedi, Ming Zhang, Yan Liu
- abstract@[open-review](#): Trajectory prediction for scenes with multiple agents and entities is a challenging problem in numerous domains such as traffic prediction, pedestrian tracking and path planning. We present a general architecture to address this challenge which models the crucial inductive biases of motion, namely, inertia, relative motion, intents and interactions. Specifically, we propose a relational model to flexibly model interactions between agents in diverse environments. Since it is well-known that human decision making is fuzzy by nature, at the core of our model lies a novel attention mechanism which models interactions by making continuous-valued (fuzzy) decisions and learning the corresponding responses. Our architecture demonstrates significant performance gains over existing state-of-the-art predictive models in diverse domains such as human crowd trajectories, US freeway traffic, NBA sports data and physics datasets. We also present ablations and augmentations to understand the decision-making process and the source of gains in our model.

[Multilabel Classification by Hierarchical Partitioning and Data-dependent Grouping](#)

- Shashanka Ubaru, Sanjeeb Dash, Arya Mazumdar, Oktay Gunluk
- abstract@[open-review](#): In modern multilabel classification problems, each data instance belongs to a small number of classes among a large set of classes. In other words, these problems involve learning very sparse binary label vectors. Moreover, in the large-scale problems, the labels typically have certain (unknown) hierarchy. In this paper we exploit the sparsity of label vectors and the hierarchical structure to embed them in low-dimensional space using label groupings. Consequently, we solve the classification problem in a much lower dimensional space and then obtain labels in the original space using an appropriately defined lifting. Our method builds on the work of (Ubaru & Mazumdar, 2017), where the idea of group testing was also explored for multilabel classification. We first present a novel data-dependent grouping approach, where we use a group construction based on a low-rank Nonnegative Matrix Factorization (NMF) of the label matrix of training instances. The construction also allows us, using recent results, to develop a fast prediction algorithm that has a logarithmic runtime in the number of labels. We then present a hierarchical partitioning approach that exploits the label hierarchy in large-scale problems to divide the large label space into smaller sub-problems, which can then be solved independently via the grouping approach. Numerical results on many benchmark datasets illustrate that, compared to other popular methods, our proposed methods achieve comparable accuracy with significantly lower computational costs.

[An Analysis of SVD for Deep Rotation Estimation](#)

- Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snavely, Angjoo Kanazawa, Afshin Rostamizadeh, Ameesh Makadia
- abstract@[open-review](#): Symmetric orthogonalization via SVD, and closely related procedures, are well-known techniques for projecting matrices onto $O(n)$ or $SO(n)$. These tools have long been used for applications in computer vision, for example optimal 3D alignment problems solved by orthogonal Procrustes, rotation averaging, or Essential matrix decomposition. Despite its utility in different settings, SVD orthogonalization as a procedure for producing rotation matrices is typically overlooked in deep learning models, where the preferences tend toward classic representations like unit quaternions, Euler angles, and axis-angle, or more recently-introduced methods. Despite the importance of 3D rotations in computer vision and robotics, a single universally effective representation is still missing. Here, we explore the viability of SVD orthogonalization for 3D rotations in neural networks. We present a theoretical analysis of SVD as used for projection onto the rotation group. Our extensive quantitative analysis shows simply replacing existing representations with the SVD orthogonalization procedure obtains state of the art performance in many deep learning applications covering both supervised and unsupervised training.

[Can the Brain Do Backpropagation? --- Exact Implementation of Backpropagation in Predictive Coding Networks](#)

- Yuhang Song, Thomas Lukasiewicz, Zhenghua Xu, Rafal Bogacz
- abstract@[open-review](#): Backpropagation (BP) has been the most successful algorithm used to train artificial neural networks. However, there are several gaps between BP and learning in biologically plausible neuronal networks of the brain (learning in the brain, or simply BL, for short), in particular, (1) it has been unclear to date, if BP can be implemented exactly via BL, (2) there is a lack of local plasticity in BP, i.e., weight updates require information that is not locally available, while BL utilizes only locally available information, and (3) there is a lack of autonomy in BP, i.e., some external control over the neural network is required (e.g., switching between prediction and learning stages requires changes to dynamics and synaptic plasticity rules), while BL works fully autonomously. Bridging such gaps, i.e., understanding how BP can be approximated by BL, has been of major interest in both neuroscience and machine learning. Despite tremendous efforts, however, no previous model has bridged the gaps at a degree of demonstrating an equivalence to BP, instead, only approximations to BP have been shown. Here, we present for the first time a framework within BL that bridges the above crucial gaps. We propose a BL model that (1) produces exactly the same updates of the neural weights as~BP, while (2) employing local plasticity, i.e., all neurons perform only local computations, done simultaneously. We then modify it to an alternative BL model that (3) also works fully autonomously. Overall, our work provides important evidence for the debate on the long-disputed question whether the brain can perform~BP.

[Manifold GPLVMs for discovering non-Euclidean latent structure in neural data](#)

- Kristopher Jensen, Ta-Chu Kao, Marco Tripodi, Guillaume Hennequin
- abstract@[open-review](#): A common problem in neuroscience is to elucidate the collective neural representations of behaviorally important variables such as head direction, spatial location, upcoming movements, or mental spatial transformations. Often, these latent variables are internal constructs not directly accessible to the experimenter. Here, we propose a new probabilistic latent variable model to simultaneously identify the latent state and the way each neuron contributes to its representation in an unsupervised way. In contrast to previous models which assume Euclidean latent spaces, we embrace the fact that latent states often belong to symmetric manifolds such as spheres, tori, or rotation groups of various dimensions. We therefore propose the manifold Gaussian process latent variable model (mGPLVM), where neural responses arise from (i) a shared latent variable living on a specific manifold, and (ii) a set of non-parametric tuning curves determining how each neuron contributes to the representation. Cross-validated comparisons of models with different topologies can be used to distinguish between candidate manifolds, and variational inference enables quantification of uncertainty. We demonstrate the validity of the approach on several synthetic datasets, as well as on calcium recordings from the ellipsoid body of *Drosophila melanogaster* and extracellular recordings from the mouse anterodorsal thalamic nucleus. These circuits are both known to encode head direction, and mGPLVM correctly recovers the ring topology expected from neural populations representing a single angular variable.

[Distributed Distillation for On-Device Learning](#)

- Ilai Bistrizt, Ariana Mann, Nicholas Bambos
- abstract@[open-review](#): On-device learning promises collaborative training of machine learning models across edge devices without the sharing of user data. In state-of-the-art on-device learning algorithms, devices communicate their model weights over a decentralized communication network. Transmitting model weights requires huge communication overhead and means only devices with identical model architectures can be included. To overcome these limitations, we introduce a distributed distillation algorithm where devices communicate and learn from soft-decision (softmax) outputs, which are inherently architecture-agnostic and scale only with the number of classes. The communicated soft-decisions are each model's outputs on a public, unlabeled reference dataset, which serves as a common vocabulary between devices. We prove that our algorithm converges with probability 1 to a

stationary point where all devices in the communication network distill the entire network's knowledge on the reference data, regardless of their local connections. Our analysis assumes smooth loss functions, which can be non-convex. Simulations support our theoretical findings and show that even a naive implementation of our algorithm significantly reduces the communication overhead while achieving an overall comparable performance to state-of-the-art, depending on the regime. By requiring little communication overhead and allowing for cross-architecture training, we remove two main obstacles to scaling on-device learning.

[COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learning](#)

- Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, Thomas Brox
- abstract@[open-review](#): Many real-world video-text tasks involve different levels of granularity, such as frames and words, clip and sentences or videos and paragraphs, each with distinct semantics. In this paper, we propose a Cooperative hierarchical Transformer (COOT) to leverage this hierarchy information and model the interactions between different levels of granularity and different modalities. The method consists of three major components: an attention-aware feature aggregation layer, which leverages the local temporal context (intra-level, e.g., within a clip), a contextual transformer to learn the interactions between low-level and high-level semantics (inter-level, e.g. clip-video, sentence-paragraph), and a cross-modal cycle-consistency loss to connect video and text. The resulting method compares favorably to the state of the art on several benchmarks while having few parameters.

[Passport-aware Normalization for Deep Model Protection](#)

- Jie Zhang, Dongdong Chen, Jing Liao, Weiming Zhang, Gang Hua, Nenghai Yu
- abstract@[open-review](#): Despite tremendous success in many application scenarios, deep learning faces serious intellectual property (IP) infringement threats. Considering the cost of designing and training a good model, infringements will significantly infringe the interests of the original model owner. Recently, many impressive works have emerged for deep model IP protection. However, they either are vulnerable to ambiguity attacks, or require changes in the target network structure by replacing its original normalization layers and hence cause significant performance drops. To this end, we propose a new passport-aware normalization formulation, which is generally applicable to most existing normalization layers and only needs to add another passport-aware branch for IP protection. This new branch is jointly trained with the target model but discarded in the inference stage. Therefore it causes no structure change in the target model. Only when the model IP is suspected to be stolen by someone, the private passport-aware branch is added back for ownership verification. Through extensive experiments, we verify its effectiveness in both image and 3D point recognition models. It is demonstrated to be robust not only to common attack techniques like fine-tuning and model compression, but also to ambiguity attacks. By further combining it with trigger-set based methods, both black-box and white-box verification can be achieved for enhanced security of deep learning models deployed in real systems.

[Sampling-Decomposable Generative Adversarial Recommender](#)

- Binbin Jin, Defu Lian, Zheng Liu, Qi Liu, Jianhui Ma, Xing Xie, Enhong Chen
- abstract@[open-review](#): Recommendation techniques are important approaches for alleviating information overload. Being often trained on implicit user feedback, many recommenders suffer from the sparsity challenge due to the lack of explicitly negative samples. The GAN-style recommenders (i.e., IRGAN) addresses the challenge by learning a generator and a discriminator adversarially, such that the generator produces increasingly difficult samples for the discriminator to accelerate optimizing the discrimination objective. However, producing samples from the generator is very time-consuming, and our empirical study shows that the discriminator performs poor in top-k item recommendation. To this end, a theoretical analysis is made for the GAN-style algorithms, showing that the generator of limit capacity is diverged from the optimal generator. This may interpret the limitation of discriminator's performance. Based on these findings, we propose a Sampling-Decomposable Generative Adversarial Recommender (SD-GAR). In the framework, the divergence between some generator and the optimum is compensated by self-normalized importance sampling; the efficiency of sample generation is improved with a sampling-decomposable generator, such that each sample can be generated in O(1) with the Vose-Alias method. Interestingly, due to decomposability of sampling, the generator can be optimized with the closed-form solutions in an alternating manner, being different from policy gradient in the GAN-style algorithms. We extensively evaluate the proposed algorithm with five real-world recommendation datasets. The results show that SD-GAR outperforms IRGAN by 12.4% and the SOTA recommender by 10% on average. Moreover, discriminator training can be 20x faster on the dataset with more than 120K items.

[Limits to Depth Efficiencies of Self-Attention](#)

- Yoav Levine, Noam Wies, Or Sharir, Hofit Bata, Amnon Shashua
- abstract@[open-review](#): Self-attention architectures, which are rapidly pushing the frontier in natural language processing, demonstrate a surprising depth-inefficient behavior: Empirical signals indicate that increasing the internal representation (network width) is just as useful as increasing the number of self-attention layers (network depth). In this paper, we theoretically study the interplay between depth and width in self-attention. We shed light on the root of the above phenomenon, and establish two distinct parameter regimes of depth efficiency and inefficiency in self-attention. We invalidate the seemingly plausible hypothesis by which widening is as effective as deepening for self-attention, and show that in fact stacking self-attention layers is so effective that it quickly saturates a capacity of the network width. Specifically, we pinpoint a ``depth threshold'' that is logarithmic in the network width: for networks of depth that is below the threshold, we establish a double-exponential depth-efficiency of the self-attention operation, while for depths over the threshold we show that depth-inefficiency kicks in. Our predictions accord with existing empirical ablations, and we further demonstrate the two depth-(in)efficiency regimes experimentally for common network depths of 6, 12, and 24. By identifying network width as a limiting factor, our analysis indicates that solutions for dramatically increasing the width can facilitate the next leap in self-attention expressivity.