Sign in to GitHub · GitHub

****

AReS and MaRS Adversarial and MMD-Minimizing Regression for SDEs

Gabriele Abbati, Philippe Wenk, Michael A. Osborne, Andreas Krause, Bernhard Schölkopf, Stefan Bauer

Stochastic differential equations are an important modeling class in many disciplines. Consequently, there exist many methods relying on various discretization and numerical integration schemes. In this paper, we propose a novel, probabilistic model for estimating the drift and diffusion given noisy observations of the underlying stochastic system. Using state-of-the-art adversarial and moment matching inference techniques, we avoid the discretization schemes of classical approaches. This leads to significant improvements in parameter accuracy and robustness given random initial guesses. On four established benchmark systems, we compare the performance of our algorithms to state-of-the-art solutions based on extended Kalman filtering and Gaussian processes.

****

Dynamic Weights in Multi-Objective Deep Reinforcement Learning

Axel Abels, Diederik Roijers, Tom Lenaerts, Ann Nowé, Denis Steckelmacher

Many real-world decision problems are characterized by multiple conflicting objectives which must be balanced based on their relative importance. In the dynamic weights setting the relative importance changes over time and specialized algorithms that deal with such change, such as a tabular Reinforcement Learning (RL) algorithm by Natarajan and Tadepalli (2005), are required. However, this earlier work is not feasible for RL settings that necessitate the use of function approximators. We generalize across weight changes and high-dimensional inputs by proposing a multi-objective Q-network whose outputs are conditioned on the relative importance of objectives and we introduce Diverse Experience Replay (DER) to counter the inherent non-stationarity of the Dynamic Weights setting. We perform an extensive experimental evaluation and compare our methods to adapted algorithms from Deep Multi-Task/Multi-Objective Reinforcement Learning and show that our proposed network in combination with DER dominates these adapted algorithms across weight change scenarios and problem domains.

****

MixHop: Higher-Order Graph Convolutional Architectures via Sparsified Neighborhood Mixing

Sami Abu-El-Haija, Bryan Perozzi, Amol Kapoor, Nazanin Alipourfard, Kristina Lerman, Hrayr Harutyunyan, Greg Ver Steeg, Aram Galstyan

Existing popular methods for semi-supervised learning with Graph Neural Networks (such as the Graph Convolutional Network) provably cannot learn a general class of neighborhood mixing relationships. To address this weakness, we propose a new model, MixHop, that can learn these relationships, including difference operators, by repeatedly mixing feature representations of neighbors at various distances. MixHop requires no additional memory or computational complexity, and outperforms on challenging baselines. In addition, we propose sparsity regularization that allows us to visualize how the network prioritizes neighborhood information across different graph datasets. Our analysis of the learned architectures reveals that neighborhood mixing varies per datasets.

****

Communication-Constrained Inference and the Role of Shared Randomness

Jayadev Acharya, Clement Canonne, Himanshu Tyagi

A central server needs to perform statistical inference based on samples that are distributed over multiple users who can each send a message of limited length to the center. We study problems of distribution learning and identity testing in this distributed inference setting and examine the role of shared randomness as a resource. We propose a general purpose simulate-and-infer strategy that uses only private-coin communication protocols and is sample-optimal for distribution learning. This general strategy turns out to be sample-optimal even for distribution testing among private-coin protocols. Interestingly, we propose a public-coin protocol that outperforms simulate-and-infer for distribution testing and is, in fact, sample-optimal. Underlying our public-coin protocol is a random hash

that when applied to the samples minimally contracts the chi-squared distance of their distribution from the uniform distribution.
****
Distributed Learning with Sublinear Communication
Jayadev Acharya, Chris De Sa, Dylan Foster, Karthik Sridharan
In distributed statistical learning, N samples are split across m machines and a learner wishes to use minimal communication to learn as well as if the examples were on a single machine. This model has received substantial interest in machine learning due to its scalability and potential for parallel speedup. However, in high-dimensional settings, where the number examples is smaller than the number of features (‘"dimension"), the speedup afforded by distributed learning may be overshadowed by the cost of communicating a single example. This paper investigates the following question: When is it possible to learn a d-dimensional model in the distributed setting with total communication sublinear in d? Starting with a negative result, we observe that for learning ℓ_1-bounded or sparse linear models, no algorithm can obtain optimal error until communication is linear in dimension. Our main result is that by slightly relaxing the standard boundedness assumptions for linear models, we can obtain distributed algorithms that enjoy optimal error with communication logarithmic in dimension. This result is based on a family of algorithms that combine mirror descent with randomized sparsification/quantization of iterates, and extends to the general stochastic convex optimization model.
****
Communication Complexity in Locally Private Distribution Estimation and Heavy Hitters
Jayadev Acharya, Ziteng Sun
We consider the problems of distribution estimation, and heavy hitter (frequency) estimation under privacy, and communication constraints. While the constraints have been studied separately, optimal schemes for one are sub-optimal for the other. We propose a sample-optimal $\eps$-locally differentially private (LDP) scheme for distribution estimation, where each user communicates one bit, and requires no public randomness. We also show that Hadamard Response, a recently proposed scheme for $\eps$-LDP distribution estimation is also utility-optimal for heavy hitters estimation. Our final result shows that unlike distribution estimation, without public randomness, any utility-optimal heavy hitter estimation algorithm must require $\Omega(\log n)$ bits of communication per user.
****
Learning Models from Data with Measurement Error: Tackling Underreporting
Roy Adams, Yuelong Ji, Xiaobin Wang, Suchi Saria
Measurement error in observational datasets can lead to systematic bias in inferences based on these datasets. As studies based on observational data are increasingly used to inform decisions with real-world impact, it is critical that we develop a robust set of techniques for analyzing and adjusting for these biases. In this paper we present a method for estimating the distribution of an outcome given a binary exposure that is subject to underreporting. Our method is based on a missing data view of the measurement error problem, where the true exposure is treated as a latent variable that is marginalized out of a joint model. We prove three different conditions under which the outcome distribution can still be identified from data containing only error-prone observations of the exposure. We demonstrate this method on synthetic data and analyze its sensitivity to near violations of the identifiability conditions. Finally, we use this method to estimate the effects of maternal smoking and heroin use during pregnancy on childhood obesity, two import problems from public health. Using the proposed method, we estimate these effects using only subject-reported drug use data and refine the range of estimates generated by a sensitivity analysis-based approach. Further, the estimates produced by our method are consistent with existing literature on both the effects of maternal smoking and the rate at which subjects underreport smoking.
****
TibGM: A Transferable and Information-Based Graphical Model Approach for Reinfor

cement Learning
Tameem Adel, Adrian Weller
One of the challenges to reinforcement learning (RL) is scalable transferability among complex tasks. Incorporating a graphical model (GM), along with the rich family of related methods, as a basis for RL frameworks provides potential to address issues such as transferability, generalisation and exploration. Here we propose a flexible GM-based RL framework which leverages efficient inference procedures to enhance generalisation and transfer power. In our proposed transferable and information-based graphical model framework ‘TibGM’, we show the equivalence between our mutual information-based objective in the GM, and an RL consolidated objective consisting of a standard reward maximisation target and a generalisation/transfer objective. In settings where there is a sparse or deceptive reward signal, our TibGM framework is flexible enough to incorporate exploration bonuses depicting intrinsic rewards. We empirically verify improved performance and exploration power.
****
PAC Learnability of Node Functions in Networked Dynamical Systems
Abhijin Adiga, Chris J Kuhlman, Madhav Marathe, S Ravi, Anil Vullikanti
We consider the PAC learnability of the local functions at the vertices of a discrete networked dynamical system, assuming that the underlying network is known. Our focus is on the learnability of threshold functions. We show that several variants of threshold functions are PAC learnable and provide tight bounds on the sample complexity. In general, when the input consists of positive and negative examples, we show that the concept class of threshold functions is not efficiently PAC learnable, unless NP = RP. Using a dynamic programming approach, we show efficient PAC learnability when the number of negative examples is small. We also present an efficient learner which is consistent with all the positive examples and at least (1-1/e) fraction of the negative examples. This algorithm is based on maximizing a submodular function under matroid constraints. By performing experiments on both synthetic and real-world networks, we study how the network structure and sample complexity influence the quality of the inferred system.
****
Static Automatic Batching In TensorFlow
Ashish Agarwal
Dynamic neural networks are becoming increasingly common, and yet it is hard to implement them efficiently. On-the-fly operation batching for such models is sub-optimal and suffers from run time overheads, while writing manually batched versions can be hard and error-prone. To address this we extend TensorFlow with pfor, a parallel-for loop optimized using static loop vectorization. With pfor, users can express computation using nested loops and conditional constructs, but get performance resembling that of a manually batched version. Benchmarks demonstrate speedups of one to two orders of magnitude on range of tasks, from jacobian computation, to Graph Neural Networks.
****
Efficient Full-Matrix Adaptive Regularization
Naman Agarwal, Brian Bullins, Xinyi Chen, Elad Hazan, Karan Singh, Cyril Zhang, Yi Zhang
Adaptive regularization methods pre-multiply a descent direction by a preconditioning matrix. Due to the large number of parameters of machine learning problems, full-matrix preconditioning methods are prohibitively expensive. We show how to modify full-matrix adaptive regularization in order to make it practical and effective. We also provide a novel theoretical analysis for adaptive regularization in non-convex optimization settings. The core of our algorithm, termed GGT, consists of the efficient computation of the inverse square root of a low-rank matrix. Our preliminary experiments show improved iteration-wise convergence rates across synthetic tasks and standard deep learning benchmarks, and that the more carefully-preconditioned steps sometimes lead to a better solution.
****
Online Control with Adversarial Disturbances
Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, Karan Singh

We study the control of linear dynamical systems with adversarial disturbances, as opposed to statistical noise. We present an efficient algorithm that achieves nearly-tight regret bounds in this setting. Our result generalizes upon previous work in two main aspects: the algorithm can accommodate adversarial noise in the dynamics, and can handle general convex costs.
****
Fair Regression: Quantitative Definitions and Reduction-Based Algorithms
Alekh Agarwal, Miroslav Dudik, Zhiwei Steven Wu
In this paper, we study the prediction of a real-valued target, such as a risk score or recidivism rate, while guaranteeing a quantitative notion of fairness with respect to a protected attribute such as gender or race. We call this class of problems fair regression. We propose general schemes for fair regression under two notions of fairness: (1) statistical parity, which asks that the prediction be statistically independent of the protected attribute, and (2) bounded group loss, which asks that the prediction error restricted to any protected group remain below some pre-determined level. While we only study these two notions of fairness, our schemes are applicable to arbitrary Lipschitz-continuous losses, and so they encompass least-squares regression, logistic regression, quantile regression, and many other tasks. Our schemes only require access to standard risk minimization algorithms (such as standard classification or least-squares regression) while providing theoretical guarantees on the optimality and fairness of the obtained solutions. In addition to analyzing theoretical properties of our schemes, we empirically demonstrate their ability to uncover fairness–accuracy frontiers on several standard datasets.
****
Learning to Generalize from Sparse and Underspecified Rewards
Rishabh Agarwal, Chen Liang, Dale Schuurmans, Mohammad Norouzi
We consider the problem of learning from sparse and underspecified rewards, where an agent receives a complex input, such as a natural language instruction, and needs to generate a complex response, such as an action sequence, while only receiving binary success-failure feedback. Such success-failure rewards are often underspecified: they do not distinguish between purposeful and accidental success. Generalization from underspecified rewards hinges on discounting spurious trajectories that attain accidental success, while learning from sparse feedback requires effective exploration. We address exploration by using a mode covering direction of KL divergence to collect a diverse set of successful trajectories, followed by a mode seeking KL divergence to train a robust policy. We propose Meta Reward Learning (MeRL) to construct an auxiliary reward function that provides more refined feedback for learning. The parameters of the auxiliary reward function are optimized with respect to the validation performance of a trained policy. The MeRL approach outperforms an alternative method for reward learning based on Bayesian Optimization, and achieves the state-of-the-art on weakly-supervised semantic parsing. It improves previous work by 1.2% and 2.4% on WikiTableQuestions and WikiSQL datasets respectively.
****
The Kernel Interaction Trick: Fast Bayesian Discovery of Pairwise Interactions in High Dimensions
Raj Agrawal, Brian Trippe, Jonathan Huggins, Tamara Broderick
Discovering interaction effects on a response of interest is a fundamental problem faced in biology, medicine, economics, and many other scientific disciplines. In theory, Bayesian methods for discovering pairwise interactions enjoy many benefits such as coherent uncertainty quantification, the ability to incorporate background knowledge, and desirable shrinkage properties. In practice, however, Bayesian methods are often computationally intractable for even moderate- dimensional problems. Our key insight is that many hierarchical models of practical interest admit a Gaussian process representation such that rather than maintaining a posterior over all $O(p^2)$ interactions, we need only maintain a vector of $O(p)$ kernel hyper-parameters. This implicit representation allows us to run Markov chain Monte Carlo (MCMC) over model hyper-parameters in time and memory linear in p per iteration. We focus on sparsity-inducing models and show on datasets with

a variety of covariate behaviors that our method: (1) reduces runtime by orders
of magnitude over naive applications of MCMC, (2) provides lower Type I and Typ
e II error relative to state-of-the-art LASSO-based approaches, and (3) offers i
mproved computational scaling in high dimensions relative to existing Bayesian a
nd LASSO-based approaches.
****
Understanding the Impact of Entropy on Policy Optimization
Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, Dale Schuurmans
Entropy regularization is commonly used to improve policy optimization in reinfo
rcement learning. It is believed to help with exploration by encouraging the sel
ection of more stochastic policies. In this work, we analyze this claim using ne
w visualizations of the optimization landscape based on randomly perturbing the
loss function. We first show that even with access to the exact gradient, policy
optimization is difficult due to the geometry of the objective function. We the
n qualitatively show that in some environments, a policy with higher entropy can
make the optimization landscape smoother, thereby connecting local optima and e
nabling the use of larger learning rates. This paper presents new tools for unde
rstanding the optimization landscape, shows that policy entropy serves as a regu
larizer, and highlights the challenge of designing general-purpose policy optimi
zation algorithms.
****
Fairwashing: the risk of rationalization
Ulrich Aivodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, A
lain Tapp
Black-box explanation is the problem of explaining how a machine learning model
– whose internal logic is hidden to the auditor and generally complex – produces
its outcomes. Current approaches for solving this problem include model explana
tion, outcome explanation as well as model inspection. While these techniques ca
n be beneficial by providing interpretability, they can be used in a negative ma
nner to perform fairwashing, which we define as promoting the false perception t
hat a machine learning model respects some ethical values. In particular, we dem
onstrate that it is possible to systematically rationalize decisions taken by an
unfair black-box model using the model explanation as well as the outcome expla
nation approaches with a given fairness metric. Our solution, LaundryML, is base
d on a regularized rule list enumeration algorithm whose objective is to search
for fair rule lists approximating an unfair black-box model. We empirically eval
uate our rationalization technique on black-box models trained on real-world dat
asets and show that one can obtain rule lists with high fidelity to the black-bo
x model while being considerably less unfair at the same time.
****
Adaptive Stochastic Natural Gradient Method for One-Shot Neural Architecture Sea
rch
Youhei Akimoto, Shinichi Shirakawa, Nozomu Yoshinari, Kento Uchida, Shota Saito,
Kouhei Nishida
High sensitivity of neural architecture search (NAS) methods against their input
such as step-size (i.e., learning rate) and search space prevents practitioners
from applying them out-of-the-box to their own problems, albeit its purpose is
to automate a part of tuning process. Aiming at a fast, robust, and widely-appli
cable NAS, we develop a generic optimization framework for NAS. We turn a couple
d optimization of connection weights and neural architecture into a differentiab
le optimization by means of stochastic relaxation. It accepts arbitrary search s
pace (widely-applicable) and enables to employ a gradient-based simultaneous opt
imization of weights and architecture (fast). We propose a stochastic natural gr
adient method with an adaptive step-size mechanism built upon our theoretical in
vestigation (robust). Despite its simplicity and no problem-dependent parameter
tuning, our method exhibited near state-of-the-art performances with low computa
tional budgets both on image classification and inpainting tasks.
****
Projections for Approximate Policy Iteration Algorithms
Riad Akrour, Joni Pajarinen, Jan Peters, Gerhard Neumann

Approximate policy iteration is a class of reinforcement learning (RL) algorithms where the policy is encoded using a function approximator and which has been especially prominent in RL with continuous action spaces. In this class of RL algorithms, ensuring increase of the policy return during policy update often requires to constrain the change in action distribution. Several approximations exist in the literature to solve this constrained policy update problem. In this paper, we propose to improve over such solutions by introducing a set of projections that transform the constrained problem into an unconstrained one which is then solved by standard gradient descent. Using these projections, we empirically demonstrate that our approach can improve the policy update solution and the control over exploration of existing approximate policy iteration algorithms.

****

Validating Causal Inference Models via Influence Functions

Ahmed Alaa, Mihaela Van Der Schaar

The problem of estimating causal effects of treatments from observational data falls beyond the realm of supervised learning {—} because counterfactual data is inaccessible, we can never observe the true causal effects. In the absence of "supervision", how can we evaluate the performance of causal inference methods? In this paper, we use influence functions {—} the functional derivatives of a loss function {—} to develop a model validation procedure that estimates the estimation error of causal inference methods. Our procedure utilizes a Taylor-like expansion to approximate the loss function of a method on a given dataset in terms of the influence functions of its loss on a "synthesized", proximal dataset with known causal effects. Under minimal regularity assumptions, we show that our procedure is consistent and efficient. Experiments on 77 benchmark datasets show that using our procedure, we can accurately predict the comparative performances of state-of-the-art causal inference methods applied to a given observational study.

****

Multi-objective training of Generative Adversarial Networks with multiple discriminators

Isabela Albuquerque, Joao Monteiro, Thang Doan, Breandan Considine, Tiago Falk, Ioannis Mitliagkas

Recent literature has demonstrated promising results for training Generative Adversarial Networks by employing a set of discriminators, in contrast to the traditional game involving one generator against a single adversary. Such methods perform single-objective optimization on some simple consolidation of the losses, e.g. an arithmetic average. In this work, we revisit the multiple-discriminator setting by framing the simultaneous minimization of losses provided by different models as a multi-objective optimization problem. Specifically, we evaluate the performance of multiple gradient descent and the hypervolume maximization algorithm on a number of different datasets. Moreover, we argue that the previously proposed methods and hypervolume maximization can all be seen as variations of multiple gradient descent in which the update direction can be computed efficiently. Our results indicate that hypervolume maximization presents a better compromise between sample quality and computational cost than previous methods.

****

Graph Element Networks: adaptive, structured computation and memory

Ferran Alet, Adarsh Keshav Jeewajee, Maria Bauza Villalonga, Alberto Rodriguez, Tomas Lozano-Perez, Leslie Kaelbling

We explore the use of graph neural networks (GNNs) to model spatial processes in which there is no a priori graphical structure. Similar to finite element analysis, we assign nodes of a GNN to spatial locations and use a computational process defined on the graph to model the relationship between an initial function defined over a space and a resulting function in the same space. We use GNNs as a computational substrate, and show that the locations of the nodes in space as well as their connectivity can be optimized to focus on the most complex parts of the space. Moreover, this representational strategy allows the learned input-output relationship to generalize over the size of the underlying space and run the same model at different levels of precision, trading computation for accuracy.

We demonstrate this method on a traditional PDE problem, a physical prediction problem from robotics, and learning to predict scene images from novel viewpoints.
****

Analogies Explained: Towards Understanding Word Embeddings
Carl Allen, Timothy Hospedales
Word embeddings generated by neural network methods such as word2vec (W2V) are well known to exhibit seemingly linear behaviour, e.g. the embeddings of analogy “woman is to queen as man is to king” approximately describe a parallelogram. This property is particularly intriguing since the embeddings are not trained to achieve it. Several explanations have been proposed, but each introduces assumptions that do not hold in practice. We derive a probabilistically grounded definition of paraphrasing that we re-interpret as word transformation, a mathematical description of “w_x is to w_y”. From these concepts we prove existence of linear relationship between W2V-type embeddings that underlie the analogical phenomenon, identifying explicit error terms.
****

Infinite Mixture Prototypes for Few-shot Learning
Kelsey Allen, Evan Shelhamer, Hanul Shin, Joshua Tenenbaum
We propose infinite mixture prototypes to adaptively represent both simple and complex data distributions for few-shot learning. Infinite mixture prototypes combine deep representation learning with Bayesian nonparametrics, representing each class by a set of clusters, unlike existing prototypical methods that represent each class by a single cluster. By inferring the number of clusters, infinite mixture prototypes interpolate between nearest neighbor and prototypical representations in a learned feature space, which improves accuracy and robustness in the few-shot regime. We show the importance of adaptive capacity for capturing complex data distributions such as super-classes (like alphabets in character recognition), with 10-25% absolute accuracy improvements over prototypical networks, while still maintaining or improving accuracy on standard few-shot learning benchmarks. By clustering labeled and unlabeled data with the same rule, infinite mixture prototypes achieve state-of-the-art semi-supervised accuracy, and can perform purely unsupervised clustering, unlike existing fully- and semi-supervised prototypical methods.
****

A Convergence Theory for Deep Learning via Over-Parameterization
Zeyuan Allen-Zhu, Yuanzhi Li, Zhao Song
Deep neural networks (DNNs) have demonstrated dominating performance in many fields; since AlexNet, networks used in practice are going wider and deeper. On the theoretical side, a long line of works have been focusing on why we can train neural networks when there is only one hidden layer. The theory of multi-layer networks remains unsettled. In this work, we prove simple algorithms such as stochastic gradient descent (SGD) can find Global Minima on the training objective of DNNs in Polynomial Time. We only make two assumptions: the inputs do not degenerate and the network is over-parameterized. The latter means the number of hidden neurons is sufficiently large: polynomial in L, the number of DNN layers and in n, the number of training samples. As concrete examples, starting from randomly initialized weights, we show that SGD attains 100% training accuracy in classification tasks, or minimizes regression loss in linear convergence speed eps   e^{-T}, with running time polynomial in n and L. Our theory applies to the widely-used but non-smooth ReLU activation, and to any smooth and possibly non-convex loss functions. In terms of network architectures, our theory at least applies to fully-connected neural networks, convolutional neural networks (CNN), and residual neural networks (ResNet).
****

Asynchronous Batch Bayesian Optimisation with Improved Local Penalisation
Ahsan Alvi, Binxin Ru, Jan-Peter Calliess, Stephen Roberts, Michael A. Osborne
Batch Bayesian optimisation (BO) has been successfully applied to hyperparameter tuning using parallel computing, but it is wasteful of resources: workers that complete jobs ahead of others are left idle. We address this problem by developi

ng an approach, Penalising Locally for Asynchronous Bayesian Optimisation on K Workers (PLAyBOOK), for asynchronous parallel BO. We demonstrate empirically the efficacy of PLAyBOOK and its variants on synthetic tasks and a real-world problem. We undertake a comparison between synchronous and asynchronous BO, and show that asynchronous BO often outperforms synchronous batch BO in both wall-clock time and sample efficiency.
****
Bounding User Contributions: A Bias-Variance Trade-off in Differential Privacy
Kareem Amin, Alex Kulesza, Andres Munoz, Sergei Vassilvtiskii
Differentially private learning algorithms protect individual participants in the training dataset by guaranteeing that their presence does not significantly change the resulting model. In order to make this promise, such algorithms need to know the maximum contribution that can be made by a single user: the more data an individual can contribute, the more noise will need to be added to protect them. While most existing analyses assume that the maximum contribution is known and fixed in advance{—}indeed, it is often assumed that each user contributes only a single example{—}we argue that in practice there is a meaningful choice to be made. On the one hand, if we allow users to contribute large amounts of data, we may end up adding excessive noise to protect a few outliers, even when the majority contribute only modestly. On the other hand, limiting users to small contributions keeps noise levels low at the cost of potentially discarding significant amounts of excess data, thus introducing bias. Here, we characterize this trade-off for an empirical risk minimization setting, showing that in general there is a “sweet spot” that depends on measurable properties of the dataset, but that there is also a concrete cost to privacy that cannot be avoided simply by collecting more data.
****
Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Value Approximation
Marco Ancona, Cengiz Oztireli, Markus Gross
The problem of explaining the behavior of deep neural networks has recently gained a lot of attention. While several attribution methods have been proposed, most come without strong theoretical foundations, which raises questions about their reliability. On the other hand, the literature on cooperative game theory suggests Shapley values as a unique way of assigning relevance scores such that certain desirable properties are satisfied. Unfortunately, the exact evaluation of Shapley values is prohibitively expensive, exponential in the number of input features. In this work, by leveraging recent results on uncertainty propagation, we propose a novel, polynomial-time approximation of Shapley values in deep neural networks. We show that our method produces significantly better approximations of Shapley values than existing state-of-the-art attribution methods.
****
Scaling Up Ordinal Embedding: A Landmark Approach
Jesse Anderton, Javed Aslam
Ordinal Embedding is the problem of placing n objects into R^d to satisfy constraints like "object a is closer to b than to c." It can accommodate data that embeddings from features or distances cannot, but is a more difficult problem. We propose a novel landmark-based method as a partial solution. At small to medium scales, we present a novel combination of existing methods with some new theoretical justification. For very large values of n optimizing over an entire embedding breaks down, so we propose a novel method which first embeds a subset of m << n objects and then embeds the remaining objects independently and in parallel. We prove a distance error bound for our method in terms of m and that it has O(dn log m) time complexity, and show empirically that it is able to produce high quality embeddings in a fraction of the time needed for any published method.
****
Sorting Out Lipschitz Function Approximation
Cem Anil, James Lucas, Roger Grosse
Training neural networks under a strict Lipschitz constraint is useful for provable adversarial robustness, generalization bounds, interpretable gradients, and

Wasserstein distance estimation. By the composition property of Lipschitz functions, it suffices to ensure that each individual affine transformation or nonlinear activation is 1-Lipschitz. The challenge is to do this while maintaining the expressive power. We identify a necessary property for such an architecture: each of the layers must preserve the gradient norm during backpropagation. Based on this, we propose to combine a gradient norm preserving activation function, GroupSort, with norm-constrained weight matrices. We show that norm-constrained GroupSort architectures are universal Lipschitz function approximators. Empirically, we show that norm-constrained GroupSort networks achieve tighter estimates of Wasserstein distance than their ReLU counterparts and can achieve provable adversarial robustness guarantees with little cost to accuracy.
****
Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data
Luigi Antelmi, Nicholas Ayache, Philippe Robert, Marco Lorenzi
Interpretable modeling of heterogeneous data channels is essential in medical applications, for example when jointly analyzing clinical scores and medical images. Variational Autoencoders (VAE) are powerful generative models that learn representations of complex data. The flexibility of VAE may come at the expense of lack of interpretability in describing the joint relationship between heterogeneous data. To tackle this problem, in this work we extend the variational framework of VAE to bring parsimony and interpretability when jointly account for latent relationships across multiple channels. In the latent space, this is achieved by constraining the variational distribution of each channel to a common target prior. Parsimonious latent representations are enforced by variational dropout. Experiments on synthetic data show that our model correctly identifies the prescribed latent dimensions and data relationships across multiple testing scenarios. When applied to imaging and clinical data, our method allows to identify the joint effect of age and pathology in describing clinical condition in a large scale clinical cohort.
****
Unsupervised Label Noise Modeling and Loss Correction
Eric Arazo, Diego Ortego, Paul Albert, Noel O’Connor, Kevin Mcguinness
Despite being robust to small amounts of label noise, convolutional neural networks trained with stochastic gradient methods have been shown to easily fit random labels. When there are a mixture of correct and mislabelled targets, networks tend to fit the former before the latter. This suggests using a suitable two-component mixture model as an unsupervised generative model of sample loss values during training to allow online estimation of the probability that a sample is mislabelled. Specifically, we propose a beta mixture to estimate this probability and correct the loss by relying on the network prediction (the so-called bootstrapping loss). We further adapt mixup augmentation to drive our approach a step further. Experiments on CIFAR-10/100 and TinyImageNet demonstrate a robustness to label noise that substantially outperforms recent state-of-the-art. Source code is available at https://git.io/fjsvE and Appendix at https://arxiv.org/abs/1904.11238.
****
Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks
Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, Ruosong Wang
Recent works have cast some light on the mystery of why deep nets fit any data and generalize despite being very overparametrized. This paper analyzes training and generalization for a simple 2-layer ReLU net with random initialization, and provides the following improvements over recent works: (i) Using a tighter characterization of training speed than recent papers, an explanation for why training a neural net with random labels leads to slower training, as originally observed in [Zhang et al. ICLR’17]. (ii) Generalization bound independent of network size, using a data-dependent complexity measure. Our measure distinguishes clearly between random labels and true labels on MNIST and CIFAR, as shown by experiments. Moreover, recent papers require sample complexity to increase (slowly) wit

h the size, while our sample complexity is completely independent of the network size. (iii) Learnability of a broad class of smooth functions by 2-layer ReLU nets trained via gradient descent. The key idea is to track dynamics of training and generalization via properties of a related kernel.
****
Distributed Weighted Matching via Randomized Composable Coresets
Sepehr Assadi, Mohammadhossein Bateni, Vahab Mirrokni
Maximum weight matching is one of the most fundamental combinatorial optimization problems with a wide range of applications in data mining and bioinformatics. Developing distributed weighted matching algorithms has been challenging due to the sequential nature of efficient algorithms for this problem. In this paper, we develop a simple distributed algorithm for the problem on general graphs with approximation guarantee of 2 + eps that (nearly) matches that of the sequential greedy algorithm. A key advantage of this algorithm is that it can be easily implemented in only two rounds of computation in modern parallel computation frameworks such as MapReduce. We also demonstrate the efficiency of our algorithm in practice on various graphs (some with half a trillion edges) by achieving objective values always close to what is achievable in the centralized setting.
****
Stochastic Gradient Push for Distributed Deep Learning
Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, Mike Rabbat
Distributed data-parallel algorithms aim to accelerate the training of deep neural networks by parallelizing the computation of large mini-batch gradient updates across multiple nodes. Approaches that synchronize nodes using exact distributed averaging (e.g., via AllReduce) are sensitive to stragglers and communication delays. The PushSum gossip algorithm is robust to these issues, but only performs approximate distributed averaging. This paper studies Stochastic Gradient Push (SGP), which combines PushSum with stochastic gradient updates. We prove that SGP converges to a stationary point of smooth, non-convex objectives at the same sub-linear rate as SGD, and that all nodes achieve consensus. We empirically validate the performance of SGP on image classification (ResNet-50, ImageNet) and machine translation (Transformer, WMT’16 En-De) workloads.
****
Bayesian Optimization of Composite Functions
Raul Astudillo, Peter Frazier
We consider optimization of composite objective functions, i.e., of the form $f(x)=g(h(x))$, where h is a black-box derivative-free expensive-to-evaluate function with vector-valued outputs, and g is a cheap-to-evaluate real-valued function. While these problems can be solved with standard Bayesian optimization, we propose a novel approach that exploits the composite structure of the objective function to substantially improve sampling efficiency. Our approach models h using a multi-output Gaussian process and chooses where to sample using the expected improvement evaluated on the implied non-Gaussian posterior on f, which we call expected improvement for composite functions (EI-CF). Although EI-CF cannot be computed in closed form, we provide a novel stochastic gradient estimator that allows its efficient maximization. We also show that our approach is asymptotically consistent, i.e., that it recovers a globally optimal solution as sampling effort grows to infinity, generalizing previous convergence results for classical expected improvement. Numerical experiments show that our approach dramatically outperforms standard Bayesian optimization benchmarks, reducing simple regret by several orders of magnitude.
****
Linear-Complexity Data-Parallel Earth Mover’s Distance Approximations
Kubilay Atasu, Thomas Mittelholzer
The Earth Mover’s Distance (EMD) is a state-of-the art metric for comparing discrete probability distributions, but its high distinguishability comes at a high cost in computational complexity. Even though linear-complexity approximation algorithms have been proposed to improve its scalability, these algorithms are either limited to vector spaces with only a few dimensions or they become ineffective when the degree of overlap between the probability distributions is high. We

propose novel approximation algorithms that overcome both of these limitations, yet still achieve linear time complexity. All our algorithms are data parallel, and therefore, we can take advantage of massively parallel computing engines, su ch as Graphics Processing Units (GPUs). On the popular text-based 20 Newsgroups dataset, the new algorithms are four orders of magnitude faster than a multi-thr eaded CPU implementation of Word Mover’s Distance and match its search accuracy. On MNIST images, the new algorithms are four orders of magnitude faster than Cu turi’s GPU implementation of the Sinkhorn’s algorithm while offering a slightly higher search accuracy.
****
Benefits and Pitfalls of the Exponential Mechanism with Applications to Hilbert Spaces and Functional PCA
Jordan Awan, Ana Kenney, Matthew Reimherr, Aleksandra Slavkovi■
The exponential mechanism is a fundamental tool of Differential Privacy (DP) due to its strong privacy guarantees and flexibility. We study its extension to set tings with summaries based on infinite dimensional outputs such as with function al data analysis, shape analysis, and nonparametric statistics. We show that the mechanism must be designed with respect to a specific base measure over the out put space, such as a Gaussian process. We provide a positive result that establi shes a Central Limit Theorem for the exponential mechanism quite broadly. We als o provide a negative result, showing that the magnitude of noise introduced for privacy is asymptotically non-negligible relative to the statistical estimation error. We develop an $\ep$-DP mechanism for functional principal component analy sis, applicable in separable Hilbert spaces, and demonstrate its performance via simulations and applications to two datasets.
****
Feature Grouping as a Stochastic Regularizer for High-Dimensional Structured Dat a
Sergul Aydore, Bertrand Thirion, Gael Varoquaux
In many applications where collecting data is expensive, for example neuroscienc e or medical imaging, the sample size is typically small compared to the feature dimension. These datasets call for intelligent regularization that exploits kno wn structure, such as correlations between the features arising from the measure ment device. However, existing structured regularizers need specially crafted so lvers, which are difficult to apply to complex models. We propose a new regulari zer specifically designed to leverage structure in the data in a way that can be applied efficiently to complex models. Our approach relies on feature grouping, using a fast clustering algorithm inside a stochastic gradient descent loop: gi ven a family of feature groupings that capture feature covariations, we randomly select these groups at each iteration. Experiments on two real-world datasets d emonstrate that the proposed approach produces models that generalize better tha n those trained with conventional regularizers, and also improves convergence sp eed, and has a linear computational cost.
****
Beyond the Chinese Restaurant and Pitman-Yor processes: Statistical Models with double power-law behavior
Fadhel Ayed, Juho Lee, Francois Caron
Bayesian nonparametric approaches, in particular the Pitman-Yor process and the associated two-parameter Chinese Restaurant process, have been successfully used in applications where the data exhibit a power-law behavior. Examples include n atural language processing, natural images or networks. There is also growing em pirical evidence suggesting that some datasets exhibit a two-regime power-law be havior: one regime for small frequencies, and a second regime, with a different exponent, for high frequencies. In this paper, we introduce a class of completel y random measures which are doubly regularly-varying. Contrary to the Pitman-Yor process, we show that when completely random measures in this class are normali zed to obtain random probability measures and associated random partitions, such partitions exhibit a double power-law behavior. We present two general construc tions and discuss in particular two models within this class: the beta prime pro cess (Broderick et al. (2015, 2018) and a novel process called generalized BFRY

process. We derive efficient Markov chain Monte Carlo algorithms to estimate the
 parameters of these models. Finally, we show that the proposed models provide a
 better fit than the Pitman-Yor process on various datasets.
****
Scalable Fair Clustering
Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, Tal
Wagner
We study the fair variant of the classic k-median problem introduced by (Chieric
hetti et al., NeurIPS 2017) in which the points are colored, and the goal is to
minimize the same average distance objective as in the standard k-median probl
em while ensuring that all clusters have an “approximately equal” number of poin
ts of each color. (Chierichetti et al., NeurIPS 2017) proposed a two-phase algor
ithm for fair k-clustering. In the first step, the pointset is partitioned int
o subsets called fairlets that satisfy the fairness requirement and approximatel
y preserve the k-median objective. In the second step, fairlets are merged into
k clusters by one of the existing k-median algorithms. The running time of this
algorithm is dominated by the first step, which takes super-quadratic time. In t
his paper, we present a practical approximate fairlet decomposition algorithm th
at runs in nearly linear time.
****
Entropic GANs meet VAEs: A Statistical Approach to Compute Sample Likelihoods in
 GANs
Yogesh Balaji, Hamed Hassani, Rama Chellappa, Soheil Feizi
Building on the success of deep learning, two modern approaches to learn a proba
bility model from the data are Generative Adversarial Networks (GANs) and Variat
ional AutoEncoders (VAEs). VAEs consider an explicit probability model for the d
ata and compute a generative distribution by maximizing a variational lower-boun
d on the log-likelihood function. GANs, however, compute a generative model by m
inimizing a distance between observed and generated probability distributions wi
thout considering an explicit model for the observed data. The lack of having ex
plicit probability models in GANs prohibits computation of sample likelihoods in
 their frameworks and limits their use in statistical inference problems. In thi
s work, we resolve this issue by constructing an explicit probability model that
 can be used to compute sample likelihood statistics in GANs. In particular, we
prove that under this probability model, a family of Wasserstein GANs with an en
tropy regularization can be viewed as a generative model that maximizes a variat
ional lower-bound on average sample log likelihoods, an approach that VAEs are b
ased on. This result makes a principled connection between two modern generative
 models, namely GANs and VAEs. In addition to the aforementioned theoretical res
ults, we compute likelihood statistics for GANs trained on Gaussian, MNIST, SVHN
, CIFAR-10 and LSUN datasets. Our numerical results validate the proposed theory
.
****
Provable Guarantees for Gradient-Based Meta-Learning
Maria-Florina Balcan, Mikhail Khodak, Ameet Talwalkar
We study the problem of meta-learning through the lens of online convex optimiza
tion, developing a meta-algorithm bridging the gap between popular gradient-base
d meta-learning and classical regularization-based multi-task transfer methods.
Our method is the first to simultaneously satisfy good sample efficiency guarant
ees in the convex setting, with generalization bounds that improve with task-sim
ilarity, while also being computationally scalable to modern deep learning archi
tectures and the many-task setting. Despite its simplicity, the algorithm matche
s, up to a constant factor, a lower bound on the performance of any such paramet
er-transfer method under natural task similarity assumptions. We use experiments
 in both convex and deep learning settings to verify and demonstrate the applica
bility of our theory.
****
Open-ended learning in symmetric zero-sum games
David Balduzzi, Marta Garnelo, Yoram Bachrach, Wojciech Czarnecki, Julien Perola
t, Max Jaderberg, Thore Graepel

Zero-sum games such as chess and poker are, abstractly, functions that evaluate
pairs of agents, for example labeling them ‘winner’ and ‘loser’. If the game is
approximately transitive, then self-play generates sequences of agents of increa
sing strength. However, nontransitive games, such as rock-paper-scissors, can ex
hibit strategic cycles, and there is no longer a clear objective – we want agent
s to increase in strength, but against whom is unclear. In this paper, we introd
uce a geometric framework for formulating agent objectives in zero-sum games, in
 order to construct adaptive sequences of objectives that yield open-ended learn
ing. The framework allows us to reason about population performance in nontransi
tive games, and enables the development of a new algorithm (rectified Nash respo
nse, PSRO_rN) that uses game-theoretic niching to construct diverse populations
of effective agents, producing a stronger set of agents than existing algorithms
. We apply PSRO_rN to two highly nontransitive resource allocation games and fin
d that PSRO_rN consistently outperforms the existing alternatives.
****
Concrete Autoencoders: Differentiable Feature Selection and Reconstruction
Muhammed Fatih Bal■n, Abubakar Abid, James Zou
We introduce the concrete autoencoder, an end-to-end differentiable method for g
lobal feature selection, which efficiently identifies a subset of the most infor
mative features and simultaneously learns a neural network to reconstruct the in
put data from the selected features. Our method is unsupervised, and is based on
 using a concrete selector layer as the encoder and using a standard neural netw
ork as the decoder. During the training phase, the temperature of the concrete s
elector layer is gradually decreased, which encourages a user-specified number o
f discrete features to be learned; during test time, the selected features can b
e used with the decoder network to reconstruct the remaining input features. We
evaluate concrete autoencoders on a variety of datasets, where they significantl
y outperform state-of-the-art methods for feature selection and data reconstruct
ion. In particular, on a large-scale gene expression dataset, the concrete autoe
ncoder selects a small subset of genes whose expression levels can be used to im
pute the expression levels of the remaining genes; in doing so, it improves on t
he current widely-used expert-curated L1000 landmark genes, potentially reducing
 measurement costs by 20%. The concrete autoencoder can be implemented by adding
 just a few lines of code to a standard autoencoder, and the code for the algori
thm and experiments is publicly available.
****
HOList: An Environment for Machine Learning of Higher Order Logic Theorem Provin
g
Kshitij Bansal, Sarah Loos, Markus Rabe, Christian Szegedy, Stewart Wilcox
We present an environment, benchmark, and deep learning driven automated theorem
 prover for higher-order logic. Higher-order interactive theorem provers enable
the formalization of arbitrary mathematical theories and thereby present an inte
resting challenge for deep learning. We provide an open-source framework based o
n the HOL Light theorem prover that can be used as a reinforcement learning envi
ronment. HOL Light comes with a broad coverage of basic mathematical theorems on
 calculus and the formal proof of the Kepler conjecture, from which we derive a
challenging benchmark for automated reasoning approaches. We also present a deep
 reinforcement learning driven automated theorem prover, DeepHOL, that gives str
ong initial results on this benchmark.
****
Structured agents for physical construction
Victor Bapst, Alvaro Sanchez-Gonzalez, Carl Doersch, Kimberly Stachenfeld, Pushm
eet Kohli, Peter Battaglia, Jessica Hamrick
Physical construction—the ability to compose objects, subject to physical dynami
cs, to serve some function—is fundamental to human intelligence. We introduce a
suite of challenging physical construction tasks inspired by how children play w
ith blocks, such as matching a target configuration, stacking blocks to connect
objects together, and creating shelter-like structures over target objects. We e
xamine how a range of deep reinforcement learning agents fare on these challenge
s, and introduce several new approaches which provide superior performance. Our

results show that agents which use structured representations (e.g., objects and scene graphs) and structured policies (e.g., object-centric actions) outperform those which use less structured representations, and generalize better beyond their training when asked to reason about larger scenes. Model-based agents which use Monte-Carlo Tree Search also outperform strictly model-free agents in our most challenging construction problems. We conclude that approaches which combine structured representations and reasoning with powerful learning are a key path toward agents that possess rich intuitive physics, scene understanding, and planning.
****

Learning to Route in Similarity Graphs
Dmitry Baranchuk, Dmitry Persiyanov, Anton Sinitsin, Artem Babenko
Recently similarity graphs became the leading paradigm for efficient nearest neighbor search, outperforming traditional tree-based and LSH-based methods. Similarity graphs perform the search via greedy routing: a query traverses the graph and in each vertex moves to the adjacent vertex that is the closest to this query. In practice, similarity graphs are often susceptible to local minima, when queries do not reach its nearest neighbors, getting stuck in suboptimal vertices. In this paper we propose to learn the routing function that overcomes local minima via incorporating information about the graph global structure. In particular, we augment the vertices of a given graph with additional representations that are learned to provide the optimal routing from the start vertex to the query nearest neighbor. By thorough experiments, we demonstrate that the proposed learnable routing successfully diminishes the local minima problem and significantly improves the overall search performance.
****

A Personalized Affective Memory Model for Improving Emotion Recognition
Pablo Barros, German Parisi, Stefan Wermter
Recent models of emotion recognition strongly rely on supervised deep learning solutions for the distinction of general emotion expressions. However, they are not reliable when recognizing online and personalized facial expressions, e.g., for person-specific affective understanding. In this paper, we present a neural model based on a conditional adversarial autoencoder to learn how to represent and edit general emotion expressions. We then propose Grow-When-Required networks as personalized affective memories to learn individualized aspects of emotional expressions. Our model achieves state-of-the-art performance on emotion recognition when evaluated on in-the-wild datasets. Furthermore, our experiments include ablation studies and neural visualizations in order to explain the behavior of our model.
****

Scale-free adaptive planning for deterministic dynamics & discounted rewards
Peter Bartlett, Victor Gabillon, Jennifer Healey, Michal Valko
We address the problem of planning in an environment with deterministic dynamics and stochastic discounted rewards under a limited numerical budget where the ranges of both rewards and noise are unknown. We introduce PlaTypOOS, an adaptive, robust, and efficient alternative to the OLOP (open-loop optimistic planning) algorithm. Whereas OLOP requires a priori knowledge of the ranges of both rewards and noise, PlaTypOOS dynamically adapts its behavior to both. This allows PlaTypOOS to be immune to two vulnerabilities of OLOP: failure when given underestimated ranges of noise and rewards and inefficiency when these are overestimated. PlaTypOOS additionally adapts to the global smoothness of the value function. PlaTypOOS acts in a provably more efficient manner vs. OLOP when OLOP is given an overestimated reward and show that in the case of no noise, PlaTypOOS learns exponentially faster.
****

Pareto Optimal Streaming Unsupervised Classification
Soumya Basu, Steven Gutstein, Brent Lance, Sanjay Shakkottai
We study an online and streaming unsupervised classification system. Our setting consists of a collection of classifiers (with unknown confusion matrices) each of which can classify one sample per unit time, and which are accessed by a stre

am of unlabeled samples. Each sample is dispatched to one or more classifiers, and depending on the labels collected from these classifiers, may be sent to other classifiers to collect additional labels. The labels are continually aggregated. Once the aggregated label has high enough accuracy (a pre-specified threshold for accuracy) or the sample is sent to all the classifiers, the now labeled sample is ejected from the system. For any given pre-specified threshold for accuracy, the objective is to sustain the maximum possible rate of arrival of new samples, such that the number of samples in memory does not grow unbounded. In this paper, we characterize the Pareto-optimal region of accuracy and arrival rate, and develop an algorithm that can operate at any point within this region. Our algorithm uses queueing-based routing and scheduling approaches combined with novel online tensor decomposition method to learn the hidden parameters, to Pareto-optimality guarantees. We finally verify our theoretical results through simulations on two ensembles formed using AlexNet, VGG, and ResNet deep image classifiers.
****
Categorical Feature Compression via Submodular Optimization
Mohammadhossein Bateni, Lin Chen, Hossein Esfandiari, Thomas Fu, Vahab Mirrokni, Afshin Rostamizadeh
In the era of big data, learning from categorical features with very large vocabularies (e.g., 28 million for the Criteo click prediction dataset) has become a practical challenge for machine learning researchers and practitioners. We design a highly-scalable vocabulary compression algorithm that seeks to maximize the mutual information between the compressed categorical feature and the target binary labels and we furthermore show that its solution is guaranteed to be within a $1-1/e \approx 63%$ factor of the global optimal solution. Although in some settings, entropy-based set functions are known to be submodular, this is not the case for the mutual information objective we consider (mutual information with respect to the target labels). To address this, we introduce a novel re-parametrization of the mutual information objective, which we prove is submodular, and also design a data structure to query the submodular function in amortized $O(\log n)$ time (where n is the input vocabulary size). Our complete algorithm is shown to operate in $O(n \log n)$ time. Additionally, we design a distributed implementation in which the query data structure is decomposed across $O(k)$ machines such that each machine only requires $O(\frac n k)$ space, while still preserving the approximation guarantee and using only logarithmic rounds of computation. We also provide analysis of simple alternative heuristic compression methods to demonstrate they cannot achieve any approximation guarantee. Using the large-scale Criteo learning task, we demonstrate better performance in retaining mutual information and also verify competitive learning performance compared to other baseline methods.
****
Noise2Self: Blind Denoising by Self-Supervision
Joshua Batson, Loic Royer
We propose a general framework for denoising high-dimensional measurements which requires no prior on the signal, no estimate of the noise, and no clean training data. The only assumption is that the noise exhibits statistical independence across different dimensions of the measurement, while the true signal exhibits some correlation. For a broad class of functions (“$\mathcal{J}$-invariant”), it is then possible to estimate the performance of a denoiser from noisy data alone. This allows us to calibrate $\mathcal{J}$-invariant versions of any parameterised denoising algorithm, from the single hyperparameter of a median filter to the millions of weights of a deep neural network. We demonstrate this on natural image and microscopy data, where we exploit noise independence between pixels, and on single-cell gene expression data, where we exploit independence between detections of individual molecules. This framework generalizes recent work on training neural nets from noisy images and on cross-validation for matrix factorization.
****
Efficient optimization of loops and limits with randomized telescoping sums

Alex Beatson, Ryan P Adams
We consider optimization problems in which the objective requires an inner loop with many steps or is the limit of a sequence of increasingly costly approximations. Meta-learning, training recurrent neural networks, and optimization of the solutions to differential equations are all examples of optimization problems with this character. In such problems, it can be expensive to compute the objective function value and its gradient, but truncating the loop or using less accurate approximations can induce biases that damage the overall solution. We propose randomized telescope (RT) gradient estimators, which represent the objective as the sum of a telescoping series and sample linear combinations of terms to provide cheap unbiased gradient estimates. We identify conditions under which RT estimators achieve optimization convergence rates independent of the length of the loop or the required accuracy of the approximation. We also derive a method for tuning RT estimators online to maximize a lower bound on the expected decrease in loss per unit of computation. We evaluate our adaptive RT estimators on a range of applications including meta-optimization of learning rates, variational inference of ODE parameters, and training an LSTM to model long sequences.
****
Recurrent Kalman Networks: Factorized Inference in High-Dimensional Deep Feature Spaces
Philipp Becker, Harit Pandya, Gregor Gebhardt, Cheng Zhao, C. James Taylor, Gerhard Neumann
In order to integrate uncertainty estimates into deep time-series modelling, Kalman Filters (KFs) (Kalman et al., 1960) have been integrated with deep learning models, however, such approaches typically rely on approximate inference tech- niques such as variational inference which makes learning more complex and often less scalable due to approximation errors. We propose a new deep approach to Kalman filtering which can be learned directly in an end-to-end manner using backpropagation without additional approximations. Our approach uses a high-dimensional factorized latent state representation for which the Kalman updates simplify to scalar operations and thus avoids hard to backpropagate, computationally heavy and potentially unstable matrix inversions. Moreover, we use locally linear dynamic models to efficiently propagate the latent state to the next time step. The resulting network architecture, which we call Recurrent Kalman Network (RKN), can be used for any time-series data, similar to a LSTM (Hochreiter & Schmidhuber, 1997) but uses an explicit representation of uncertainty. As shown by our experiments, the RKN obtains much more accurate uncertainty estimates than an LSTM or Gated Recurrent Units (GRUs) (Cho et al., 2014) while also showing a slightly improved prediction performance and outperforms various recent generative models on an image imputation task.
****
Switching Linear Dynamics for Variational Bayes Filtering
Philip Becker-Ehmck, Jan Peters, Patrick Van Der Smagt
System identification of complex and nonlinear systems is a central problem for model predictive control and model-based reinforcement learning. Despite their complexity, such systems can often be approximated well by a set of linear dynamical systems if broken into appropriate subsequences. This mechanism not only helps us find good approximations of dynamics, but also gives us deeper insight into the underlying system. Leveraging Bayesian inference, Variational Autoencoders and Concrete relaxations, we show how to learn a richer and more meaningful state space, e.g. encoding joint constraints and collisions with walls in a maze, from partial and high-dimensional observations. This representation translates into a gain of accuracy of learned dynamics showcased on various simulated tasks.
****
Active Learning for Probabilistic Structured Prediction of Cuts and Matchings
Sima Behpour, Anqi Liu, Brian Ziebart
Active learning methods, like uncertainty sampling, combined with probabilistic prediction techniques have achieved success in various problems like image classification and text classification. For more complex multivariate prediction tasks, the relationships between labels play an important role in designing structur

ed classifiers with better performance. However, computational time complexity limits prevalent probabilistic methods from effectively supporting active learning. Specifically, while non-probabilistic methods based on structured support vector ma-chines can be tractably applied to predicting cuts and bipartite matchings, conditional random fields are intractable for these structures. We propose an adversarial approach for active learning with structured prediction domains that is tractable for cuts and matching. We evaluate this approach algorithmically in two important structured prediction problems: multi-label classification and object tracking in videos. We demonstrate better accuracy and computational efficiency for our proposed method.
****

Invertible Residual Networks
Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, Joern-Henrik Jacobsen
We show that standard ResNet architectures can be made invertible, allowing the same model to be used for classification, density estimation, and generation. Typically, enforcing invertibility requires partitioning dimensions or restricting network architectures. In contrast, our approach only requires adding a simple normalization step during training, already available in standard frameworks. Invertible ResNets define a generative model which can be trained by maximum likelihood on unlabeled data. To compute likelihoods, we introduce a tractable approximation to the Jacobian log-determinant of a residual block. Our empirical evaluation shows that invertible ResNets perform competitively with both state-of-the-art image classifiers and flow-based generative models, something that has not been previously achieved with a single architecture.
****

Greedy Layerwise Learning Can Scale To ImageNet
Eugene Belilovsky, Michael Eickenberg, Edouard Oyallon
Shallow supervised 1-hidden layer neural networks have a number of favorable properties that make them easier to interpret, analyze, and optimize than their deep counterparts, but lack their representational power. Here we use 1-hidden layer learning problems to sequentially build deep networks layer by layer, which can inherit properties from shallow networks. Contrary to previous approaches using shallow networks, we focus on problems where deep learning is reported as critical for success. We thus study CNNs on image classification tasks using the large-scale ImageNet dataset and the CIFAR-10 dataset. Using a simple set of ideas for architecture and training we find that solving sequential 1-hidden-layer auxiliary problems lead to a CNN that exceeds AlexNet performance on ImageNet. Extending this training methodology to construct individual layers by solving 2-and-3-hidden layer auxiliary problems, we obtain an 11-layer network that exceeds several members of the VGG model family on ImageNet, and can train a VGG-11 model to the same accuracy as end-to-end learning. To our knowledge, this is the first competitive alternative to end-to-end training of CNNs that can scale to ImageNet. We illustrate several interesting properties of these models and conduct a range of experiments to study the properties this training induces on the intermediate layers.
****

Overcoming Multi-model Forgetting
Yassine Benyahia, Kaicheng Yu, Kamil Bennani Smires, Martin Jaggi, Anthony C. Davison, Mathieu Salzmann, Claudiu Musat
We identify a phenomenon, which we refer to as multi-model forgetting, that occurs when sequentially training multiple deep networks with partially-shared parameters; the performance of previously-trained models degrades as one optimizes a subsequent one, due to the overwriting of shared parameters. To overcome this, we introduce a statistically-justified weight plasticity loss that regularizes the learning of a model’s shared parameters according to their importance for the previous models, and demonstrate its effectiveness when training two models sequentially and for neural architecture search. Adding weight plasticity in neural architecture search preserves the best models to the end of the search and yields improved results in both natural language processing and computer vision tasks

.
****
Optimal Kronecker-Sum Approximation of Real Time Recurrent Learning
Frederik Benzing, Marcelo Matheus Gauy, Asier Mujika, Anders Martinsson, Angelik
a Steger
One of the central goals of Recurrent Neural Networks (RNNs) is to learn long-te
rm dependencies in sequential data. Nevertheless, the most popular training meth
od, Truncated Backpropagation through Time (TBPTT), categorically forbids learni
ng dependencies beyond the truncation horizon. In contrast, the online training
algorithm Real Time Recurrent Learning (RTRL) provides untruncated gradients, wi
th the disadvantage of impractically large computational costs. Recently publish
ed approaches reduce these costs by providing noisy approximations of RTRL. We p
resent a new approximation algorithm of RTRL, Optimal Kronecker-Sum Approximatio
n (OK). We prove that OK is optimal for a class of approximations of RTRL, which
 includes all approaches published so far. Additionally, we show that OK has emp
irically negligible noise: Unlike previous algorithms it matches TBPTT in a real
 world task (character-level Penn TreeBank) and can exploit online parameter upd
ates to outperform TBPTT in a synthetic string memorization task. Code available
 at GitHub.
****
Adversarially Learned Representations for Information Obfuscation and Inference
Martin Bertran, Natalia Martinez, Afroditi Papadaki, Qiang Qiu, Miguel Rodrigues
, Galen Reeves, Guillermo Sapiro
Data collection and sharing are pervasive aspects of modern society. This proces
s can either be voluntary, as in the case of a person taking a facial image to u
nlock his/her phone, or incidental, such as traffic cameras collecting videos on
 pedestrians. An undesirable side effect of these processes is that shared data
can carry information about attributes that users might consider as sensitive, e
ven when such information is of limited use for the task. It is therefore desira
ble for both data collectors and users to design procedures that minimize sensit
ive information leakage. Balancing the competing objectives of providing meaning
ful individualized service levels and inference while obfuscating sensitive info
rmation is still an open problem. In this work, we take an information theoretic
 approach that is implemented as an unconstrained adversarial game between Deep
Neural Networks in a principled, data-driven manner. This approach enables us to
 learn domain-preserving stochastic transformations that maintain performance on
 existing algorithms while minimizing sensitive information leakage.
****
Bandit Multiclass Linear Classification: Efficient Algorithms for the Separable
Case
Alina Beygelzimer, David Pal, Balazs Szorenyi, Devanathan Thiruvenkatachari, Che
n-Yu Wei, Chicheng Zhang
We study the problem of efficient online multiclass linear classification with b
andit feedback, where all examples belong to one of K classes and lie in the $
d$-dimensional Euclidean space. Previous works have left open the challenge of d
esigning efficient algorithms with finite mistake bounds when the data is linear
ly separable by a margin γ. In this work, we take a first step towards th
is problem. We consider two notions of linear separability: strong and weak. 1.
Under the strong linear separability condition, we design an efficient algorithm
 that achieves a near-optimal mistake bound of $O\left(\frac{K}{\gamma^2} \right
)$. 2. Under the more challenging weak linear separability condition, we design
an efficient algorithm with a mistake bound of $2^{\widetilde{O}(\min(K \log^2 \
frac{1}{\gamma}, \sqrt{\frac{1}{\gamma}} \log K))}$. Our algorithm is based on k
ernel Perceptron, which is inspired by the work of Klivans & Servedio (2008) on
improperly learning intersection of halfspaces.
****
Analyzing Federated Learning through an Adversarial Lens
Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, Seraphin Calo
Federated learning distributes model training among a multitude of agents, who,
guided by privacy concerns, perform training using their local data but share on

ly model parameter updates, for iterative aggregation at the server to train an
overall global model. In this work, we explore how the federated learning settin
g gives rise to a new threat, namely model poisoning, which differs from traditi
onal data poisoning. Model poisoning is carried out by an adversary controlling
a small number of malicious agents (usually 1) with the aim of causing the globa
l model to misclassify a set of chosen inputs with high con■dence. We explore a
number of strategies to carry out this attack on deep neural networks, starting
with targeted model poisoning using a simple boosting of the malicious agent’s u
pdate to overcome the effects of other agents. We also propose two critical noti
ons of stealth to detect malicious updates. We bypass these by including them in
 the adversarial objective to carry out stealthy model poisoning. We improve its
 stealth with the use of an alternating minimization strategy which alternately
optimizes for stealth and the adversarial objective. We also empirically demonst
rate that Byzantine-resilient aggregation strategies are not robust to our attac
ks. Our results indicate that highly constrained adversaries can carry out model
 poisoning attacks while maintaining stealth, thus highlighting the vulnerabilit
y of the federated learning setting and the need to develop effective defense st
rategies.
****
Optimal Continuous DR-Submodular Maximization and Applications to Provable Mean
Field Inference
Yatao Bian, Joachim Buhmann, Andreas Krause
Mean field inference for discrete graphical models is generally a highly nonconv
ex problem, which also holds for the class of probabilistic log-submodular model
s. Existing optimization methods, e.g., coordinate ascent algorithms, typically
only find local optima. In this work we propose provable mean filed methods for
probabilistic log-submodular models and its posterior agreement (PA) with strong
 approximation guarantees. The main algorithmic technique is a new Double Greedy
 scheme, termed DR-DoubleGreedy, for continuous DR-submodular maximization with
box-constraints. It is a one-pass algorithm with linear time complexity, reachin
g the optimal 1/2 approximation ratio, which may be of independent interest. We
validate the superior performance of our algorithms against baselines on both sy
nthetic and real-world datasets.
****
More Efficient Off-Policy Evaluation through Regularized Targeted Learning
Aurelien Bibaut, Ivana Malenica, Nikos Vlassis, Mark Van Der Laan
We study the problem of off-policy evaluation (OPE) in Reinforcement Learning (R
L), where the aim is to estimate the performance of a new policy given historica
l data that may have been generated by a different policy, or policies. In parti
cular, we introduce a novel doubly-robust estimator for the OPE problem in RL, b
ased on the Targeted Maximum Likelihood Estimation principle from the statistica
l causal inference literature. We also introduce several variance reduction tech
niques that lead to impressive performance gains in off-policy evaluation. We sh
ow empirically that our estimator uniformly wins over existing off-policy evalua
tion methods across multiple RL environments and various levels of model misspec
ification. Finally, we further the existing theoretical analysis of estimators f
or the RL off-policy estimation problem by showing their $O_P(1/\sqrt{n})$ rate
of convergence and characterizing their asymptotic distribution.
****
A Kernel Perspective for Regularizing Deep Neural Networks
Alberto Bietti, Grégoire Mialon, Dexiong Chen, Julien Mairal
We propose a new point of view for regularizing deep neural networks by using th
e norm of a reproducing kernel Hilbert space (RKHS). Even though this norm canno
t be computed, it admits upper and lower approximations leading to various pract
ical strategies. Specifically, this perspective (i) provides a common umbrella f
or many existing regularization principles, including spectral norm and gradient
 penalties, or adversarial training, (ii) leads to new effective regularization
penalties, and (iii) suggests hybrid strategies combining lower and upper bounds
 to get better approximations of the RKHS norm. We experimentally show this appr
oach to be effective when learning on small datasets, or to obtain adversarially

robust models.
****
Rethinking Lossy Compression: The Rate-Distortion-Perception Tradeoff
Yochai Blau, Tomer Michaeli
Lossy compression algorithms are typically designed and analyzed through the len
s of Shannon’s rate-distortion theory, where the goal is to achieve the lowest p
ossible distortion (e.g., low MSE or high SSIM) at any given bit rate. However,
in recent years, it has become increasingly accepted that "low distortion" is no
t a synonym for "high perceptual quality", and in fact optimization of one often
comes at the expense of the other. In light of this understanding, it is natura
l to seek for a generalization of rate-distortion theory which takes perceptual
quality into account. In this paper, we adopt the mathematical definition of per
ceptual quality recently proposed by Blau & Michaeli (2018), and use it to study
the three-way tradeoff between rate, distortion, and perception. We show that r
estricting the perceptual quality to be high, generally leads to an elevation of
the rate-distortion curve, thus necessitating a sacrifice in either rate or dis
tortion. We prove several fundamental properties of this triple-tradeoff, calcul
ate it in closed form for a Bernoulli source, and illustrate it visually on a to
y MNIST example.
****
Correlated bandits or: How to minimize mean-squared error online
Vinay Praneeth Boda, Prashanth L.A.
While the objective in traditional multi-armed bandit problems is to find the ar
m with the highest mean, in many settings, finding an arm that best captures inf
ormation about other arms is of interest. This objective, however, requires lear
ning the underlying correlation structure and not just the means. Sensors placem
ent for industrial surveillance and cellular network monitoring are a few applic
ations, where the underlying correlation structure plays an important role. Moti
vated by such applications, we formulate the correlated bandit problem, where th
e objective is to find the arm with the lowest mean-squared error (MSE) in estim
ating all the arms. To this end, we derive first an MSE estimator based on sampl
e variances/covariances and show that our estimator exponentially concentrates a
round the true MSE. Under a best-arm identification framework, we propose a succ
essive rejects type algorithm and provide bounds on the probability of error in
identifying the best arm. Using minimax theory, we also derive fundamental perfo
rmance limits for the correlated bandit problem.
****
Adversarial Attacks on Node Embeddings via Graph Poisoning
Aleksandar Bojchevski, Stephan Günnemann
The goal of network representation learning is to learn low-dimensional node emb
eddings that capture the graph structure and are useful for solving downstream t
asks. However, despite the proliferation of such methods, there is currently no
study of their robustness to adversarial attacks. We provide the first adversari
al vulnerability analysis on the widely used family of methods based on random w
alks. We derive efficient adversarial perturbations that poison the network stru
cture and have a negative effect on both the quality of the embeddings and the d
ownstream tasks. We further show that our attacks are transferable since they ge
neralize to many models and are successful even when the attacker is restricted.
****
Online Variance Reduction with Mixtures
Zalán Borsos, Sebastian Curi, Kfir Yehuda Levy, Andreas Krause
Adaptive importance sampling for stochastic optimization is a promising approach
that offers improved convergence through variance reduction. In this work, we p
ropose a new framework for variance reduction that enables the use of mixtures o
ver predefined sampling distributions, which can naturally encode prior knowledg
e about the data. While these sampling distributions are fixed, the mixture weig
hts are adapted during the optimization process. We propose VRM, a novel and eff
icient adaptive scheme that asymptotically recovers the best mixture weights in
hindsight and can also accommodate sampling distributions over sets of points. W
e empirically demonstrate the versatility of VRM in a range of applications.

****

Compositional Fairness Constraints for Graph Embeddings

Avishek Bose, William Hamilton

Learning high-quality node embeddings is a key building block for machine learning models that operate on graph data, such as social networks and recommender systems. However, existing graph embedding techniques are unable to cope with fairness constraints, e.g., ensuring that the learned representations do not correlate with certain attributes, such as age or gender. Here, we introduce an adversarial framework to enforce fairness constraints on graph embeddings. Our approach is compositional—meaning that it can flexibly accommodate different combinations of fairness constraints during inference. For instance, in the context of social recommendations, our framework would allow one user to request that their recommendations are invariant to both their age and gender, while also allowing another user to request invariance to just their age. Experiments on standard knowledge graph and recommender system benchmarks highlight the utility of our proposed framework.

****

Unreproducible Research is Reproducible

Xavier Bouthillier, César Laurent, Pascal Vincent

The apparent contradiction in the title is a wordplay on the different meanings attributed to the word reproducible across different scientific fields. What we imply is that unreproducible findings can be built upon reproducible methods. Without denying the importance of facilitating the reproduction of methods, we deem important to reassert that reproduction of findings is a fundamental step of the scientific inquiry. We argue that the commendable quest towards easy deterministic reproducibility of methods and numerical results should not have us forget the even more important necessity of ensuring the reproducibility of empirical findings and conclusions by properly accounting for essential sources of variations. We provide experiments to exemplify the brittleness of current common practice in the evaluation of models in the field of deep learning, showing that even if the results could be reproduced, a slightly different experiment would not support the findings. We hope to help clarify the distinction between exploratory and empirical research in the field of deep learning and believe more energy should be devoted to proper empirical research in our community. This work is an attempt to promote the use of more rigorous and diversified methodologies. It is not an attempt to impose a new methodology and it is not a critique on the nature of exploratory research.

****

Blended Conditonal Gradients

Gábor Braun, Sebastian Pokutta, Dan Tu, Stephen Wright

We present a blended conditional gradient approach for minimizing a smooth convex function over a polytope P, combining the Frank{–}Wolfe algorithm (also called conditional gradient) with gradient-based steps, different from away steps and pairwise steps, but still achieving linear convergence for strongly convex functions, along with good practical performance. Our approach retains all favorable properties of conditional gradient algorithms, notably avoidance of projections onto P and maintenance of iterates as sparse convex combinations of a limited number of extreme points of P. The algorithm is lazy, making use of inexpensive inexact solutions of the linear programming subproblem that characterizes the conditional gradient approach. It decreases measures of optimality (primal and dual gaps) rapidly, both in the number of iterations and in wall-clock time, outperforming even the lazy conditional gradient algorithms of Braun et al. 2017. We also present a streamlined version of the algorithm that applies when P is the probability simplex.

****

Coresets for Ordered Weighted Clustering

Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, Xuan Wu

We design coresets for Ordered k-Median, a generalization of classical clustering problems such as k-Median and k-Center. Its objective function is defined via the Ordered Weighted Averaging (OWA) paradigm of Yager (1988), where data points

are weighted according to a predefined weight vector, but in order of their contribution to the objective (distance from the centers). A powerful data-reduction technique, called a coreset, is to summarize a point set X in $\mathbb{R}^d$ into a small (weighted) point set X', such that for every set of k potential centers, the objective value of the coreset X' approximates that of X within factor $1\pm \epsilon$. When there are multiple objectives (weights), the above standard coreset might have limited usefulness, whereas in a simultaneous coreset, the above approximation holds for all weights (in addition to all centers). Our main result is a construction of a simultaneous coreset of size $O_{\epsilon, d}(k^2 \log^2 |X|)$ for Ordered k-Median. We validate our algorithm on a real geographical data set, and we find our coreset leads to a massive speedup of clustering computations, while maintaining high accuracy for a range of weights.
****
Target Tracking for Contextual Bandits: Application to Demand Side Management
Margaux Brégère, Pierre Gaillard, Yannig Goude, Gilles Stoltz
We propose a contextual-bandit approach for demand side management by offering price incentives. More precisely, a target mean consumption is set at each round and the mean consumption is modeled as a complex function of the distribution of prices sent and of some contextual variables such as the temperature, weather, and so on. The performance of our strategies is measured in quadratic losses through a regret criterion. We offer $T^{2/3}$ upper bounds on this regret (up to poly-logarithmic terms)—and even faster rates under stronger assumptions—for strategies inspired by standard strategies for contextual bandits (like LinUCB, see Li et al., 2010). Simulations on a real data set gathered by UK Power Networks, in which price incentives were offered, show that our strategies are effective and may indeed manage demand response by suitably picking the price levels.
****
Active Manifolds: A non-linear analogue to Active Subspaces
Robert Bridges, Anthony Gruber, Christopher Felder, Miki Verma, Chelsey Hoff
We present an approach to analyze $C^1(\mathbb{R}^m)$ functions that addresses limitations present in the Active Subspaces (AS) method of Constantine et al. (2014; 2015). Under appropriate hypotheses, our Active Manifolds (AM) method identifies a 1-D curve in the domain (the active manifold) on which nearly all values of the unknown function are attained, which can be exploited for approximation or analysis, especially when m is large (high-dimensional input space). We provide theorems justifying our AM technique and an algorithm permitting functional approximation and sensitivity analysis. Using accessible, low-dimensional functions as initial examples, we show AM reduces approximation error by an order of magnitude compared to AS, at the expense of more computation. Following this, we revisit the sensitivity analysis by Glaws et al. (2017), who apply AS to analyze a magnetohydrodynamic power generator model, and compare the performance of AM on the same data. Our analysis provides detailed information not captured by AS, exhibiting the influence of each parameter individually along an active manifold. Overall, AM represents a novel technique for analyzing functional models with benefits including: reducing m-dimensional analysis to a 1-D analogue, permitting more accurate regression than AS (at more computational expense), enabling more informative sensitivity analysis, and granting accessible visualizations (2-D plots) of parameter sensitivity along the AM.
****
Conditioning by adaptive sampling for robust design
David Brookes, Hahnbeom Park, Jennifer Listgarten
We present a method for design problems wherein the goal is to maximize or specify the value of one or more properties of interest (e.g. maximizing the fluorescence of a protein). We assume access to black box, stochastic “oracle" predictive functions, each of which maps from design space to a distribution over properties of interest. Because many state-of-the-art predictive models are known to suffer from pathologies, especially for data far from the training distribution, the problem becomes different from directly optimizing the oracles. Herein, we propose a method to solve this problem that uses model-based adaptive sampling to estimate a distribution over the design space, conditioned on the desired proper

ties.
****
Extrapolating Beyond Suboptimal Demonstrations via Inverse Reinforcement Learning from Observations
Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, Scott Niekum
A critical flaw of existing inverse reinforcement learning (IRL) methods is their inability to significantly outperform the demonstrator. This is because IRL typically seeks a reward function that makes the demonstrator appear near-optimal, rather than inferring the underlying intentions of the demonstrator that may have been poorly executed in practice. In this paper, we introduce a novel reward-learning-from-observation algorithm, Trajectory-ranked Reward EXtrapolation (T-REX), that extrapolates beyond a set of (approximately) ranked demonstrations in order to infer high-quality reward functions from a set of potentially poor demonstrations. When combined with deep reinforcement learning, T-REX outperforms state-of-the-art imitation learning and IRL methods on multiple Atari and MuJoCo benchmark tasks and achieves performance that is often more than twice the performance of the best demonstration. We also demonstrate that T-REX is robust to ranking noise and can accurately extrapolate intention by simply watching a learner noisily improve at a task over time.
****
Deep Counterfactual Regret Minimization
Noam Brown, Adam Lerer, Sam Gross, Tuomas Sandholm
Counterfactual Regret Minimization (CFR) is the leading algorithm for solving large imperfect-information games. It converges to an equilibrium by iteratively traversing the game tree. In order to deal with extremely large games, abstraction is typically applied before running CFR. The abstracted game is solved with tabular CFR, and its solution is mapped back to the full game. This process can be problematic because aspects of abstraction are often manual and domain specific, abstraction algorithms may miss important strategic nuances of the game, and there is a chicken-and-egg problem because determining a good abstraction requires knowledge of the equilibrium of the game. This paper introduces Deep Counterfactual Regret Minimization, a form of CFR that obviates the need for abstraction by instead using deep neural networks to approximate the behavior of CFR in the full game. We show that Deep CFR is principled and achieves strong performance in large poker games. This is the first non-tabular variant of CFR to be successful in large games.
****
Understanding the Origins of Bias in Word Embeddings
Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, Richard Zemel
Popular word embedding algorithms exhibit stereotypical biases, such as gender bias. The widespread use of these algorithms in machine learning systems can amplify stereotypes in important contexts. Although some methods have been developed to mitigate this problem, how word embedding biases arise during training is poorly understood. In this work we develop a technique to address this question. Given a word embedding, our method reveals how perturbing the training corpus would affect the resulting embedding bias. By tracing the origins of word embedding bias back to the original training documents, one can identify subsets of documents whose removal would most reduce bias. We demonstrate our methodology on Wikipedia and New York Times corpora, and find it to be very accurate.
****
Low Latency Privacy Preserving Inference
Alon Brutzkus, Ran Gilad-Bachrach, Oren Elisha
When applying machine learning to sensitive data, one has to find a balance between accuracy, information security, and computational-complexity. Recent studies combined Homomorphic Encryption with neural networks to make inferences while protecting against information leakage. However, these methods are limited by the width and depth of neural networks that can be used (and hence the accuracy) and exhibit high latency even for relatively simple networks. In this study we provide two solutions that address these limitations. In the first solution, we present more than 10\times improvement in latency and enable inference on wider net

works compared to prior attempts with the same level of security. The improved performance is achieved by novel methods to represent the data during the computation. In the second solution, we apply the method of transfer learning to provide private inference services using deep networks with latency of \sim0.16 seconds. We demonstrate the efficacy of our methods on several computer vision tasks.
****

Why do Larger Models Generalize Better? A Theoretical Perspective via the XOR Problem

Alon Brutzkus, Amir Globerson

Empirical evidence suggests that neural networks with ReLU activations generalize better with over-parameterization. However, there is currently no theoretical analysis that explains this observation. In this work, we provide theoretical and empirical evidence that, in certain cases, overparameterized convolutional networks generalize better than small networks because of an interplay between weight clustering and feature exploration at initialization. We demonstrate this theoretically for a 3-layer convolutional neural network with max-pooling, in a novel setting which extends the XOR problem. We show that this interplay implies that with overparamterization, gradient descent converges to global minima with better generalization performance compared to global minima of small networks. Empirically, we demonstrate these phenomena for a 3-layer convolutional neural network in the MNIST task.
****

Adversarial examples from computational constraints

Sebastien Bubeck, Yin Tat Lee, Eric Price, Ilya Razenshteyn

Why are classifiers in high dimension vulnerable to “adversarial” perturbations? We show that it is likely not due to information theoretic limitations, but rather it could be due to computational constraints. First we prove that, for a broad set of classification tasks, the mere existence of a robust classifier implies that it can be found by a possibly exponential-time algorithm with relatively few training examples. Then we give two particular classification tasks where learning a robust classifier is computationally intractable. More precisely we construct two binary classifications task in high dimensional space which are (i) information theoretically easy to learn robustly for large perturbations, (ii) efficiently learnable (non-robustly) by a simple linear separator, (iii) yet are not efficiently robustly learnable, even for small perturbations. Specifically, for the first task hardness holds for any efficient algorithm in the statistical query (SQ) model, while for the second task we rule out any efficient algorithm under a cryptographic assumption. These examples give an exponential separation between classical learning and robust learning in the statistical query model or under a cryptographic assumption. It suggests that adversarial examples may be an unavoidable byproduct of computational limitations of learning algorithms.
****

Self-similar Epochs: Value in arrangement

Eliav Buchnik, Edith Cohen, Avinatan Hasidim, Yossi Matias

Optimization of machine learning models is commonly performed through stochastic gradient updates on randomly ordered training examples. This practice means that each fraction of an epoch comprises an independent random sample of the training data that may not preserve informative structure present in the full data. We hypothesize that the training can be more effective with self-similar arrangements that potentially allow each epoch to provide benefits of multiple ones. We study this for “matrix factorization” – the common task of learning metric embeddings of entities such as queries, videos, or words from example pairwise associations. We construct arrangements that preserve the weighted Jaccard similarities of rows and columns and experimentally observe training acceleration of 3%-37% on synthetic and recommendation datasets. Principled arrangements of training examples emerge as a novel and potentially powerful enhancement to SGD that merits further exploration.
****

Learning Generative Models across Incomparable Spaces

Charlotte Bunne, David Alvarez-Melis, Andreas Krause, Stefanie Jegelka

Generative Adversarial Networks have shown remarkable success in learning a distribution that faithfully recovers a reference distribution in its entirety. However, in some cases, we may want to only learn some aspects (e.g., cluster or manifold structure), while modifying others (e.g., style, orientation or dimension). In this work, we propose an approach to learn generative models across such incomparable spaces, and demonstrate how to steer the learned distribution towards target properties. A key component of our model is the Gromov-Wasserstein distance, a notion of discrepancy that compares distributions relationally rather than absolutely. While this framework subsumes current generative models in identically reproducing distributions, its inherent flexibility allows application to tasks in manifold learning, relational learning and cross-domain learning.
****
Rates of Convergence for Sparse Variational Gaussian Process Regression
David Burt, Carl Edward Rasmussen, Mark Van Der Wilk
Excellent variational approximations to Gaussian process posteriors have been developed which avoid the $\mathcal{O}\left(N^3\right)$ scaling with dataset size N. They reduce the computational cost to $\mathcal{O}\left(NM^2\right)$, with $M\ll N$ the number of inducing variables, which summarise the process. While the computational cost seems to be linear in N, the true complexity of the algorithm depends on how M must increase to ensure a certain quality of approximation. We show that with high probability the KL divergence can be made arbitrarily small by growing M more slowly than N. A particular case is that for regression with normally distributed inputs in D-dimensions with the Squared Exponential kernel, $M=\mathcal{O}(\log^D N)$ suffices. Our results show that as datasets grow, Gaussian process posteriors can be approximated cheaply, and provide a concrete rule for how to increase M in continual learning scenarios.
****
What is the Effect of Importance Weighting in Deep Learning?
Jonathon Byrd, Zachary Lipton
Importance-weighted risk minimization is a key ingredient in many machine learning algorithms for causal inference, domain adaptation, class imbalance, and off-policy reinforcement learning. While the effect of importance weighting is well-characterized for low-capacity misspecified models, little is known about how it impacts over-parameterized, deep neural networks. This work is inspired by recent theoretical results showing that on (linearly) separable data, deep linear networks optimized by SGD learn weight-agnostic solutions, prompting us to ask, for realistic deep networks, for which many practical datasets are separable, what is the effect of importance weighting? We present the surprising finding that while importance weighting impacts models early in training, its effect diminishes over successive epochs. Moreover, while L2 regularization and batch normalization (but not dropout), restore some of the impact of importance weighting, they express the effect via (seemingly) the wrong abstraction: why should practitioners tweak the L2 regularization, and by how much, to produce the correct weighting effect? Our experiments confirm these findings across a range of architectures and datasets.
****
A Quantitative Analysis of the Effect of Batch Normalization on Gradient Descent
Yongqiang Cai, Qianxiao Li, Zuowei Shen
Despite its empirical success and recent theoretical progress, there generally lacks a quantitative analysis of the effect of batch normalization (BN) on the convergence and stability of gradient descent. In this paper, we provide such an analysis on the simple problem of ordinary least squares (OLS), where the precise dynamical properties of gradient descent (GD) is completely known, thus allowing us to isolate and compare the additional effects of BN. More precisely, we show that unlike GD, gradient descent with BN (BNGD) converges for arbitrary learning rates for the weights, and the convergence remains linear under mild conditions. Moreover, we quantify two different sources of acceleration of BNGD over GD – one due to over-parameterization which improves the effective condition number and another due having a large range of learning rates giving rise to fast descent. These phenomena set BNGD apart from GD and could account for much of its ro

bustness properties. These findings are confirmed quantitatively by numerical experiments, which further show that many of the uncovered properties of BNGD in OLS are also observed qualitatively in more complex supervised learning problems.
****
Accelerated Linear Convergence of Stochastic Momentum Methods in Wasserstein Distances
Bugra Can, Mert Gurbuzbalaban, Lingjiong Zhu
Momentum methods such as Polyak's heavy ball (HB) method, Nesterov's accelerated gradient (AG) as well as accelerated projected gradient (APG) method have been commonly used in machine learning practice, but their performance is quite sensitive to noise in the gradients. We study these methods under a first-order stochastic oracle model where noisy estimates of the gradients are available. For strongly convex problems, we show that the distribution of the iterates of AG converges with the accelerated $O(\sqrt{\kappa}\log(1/\varepsilon))$ linear rate to a ball of radius ε centered at a unique invariant distribution in the 1-Wasserstein metric where κ is the condition number as long as the noise variance is smaller than an explicit upper bound we can provide. Our analysis also certifies linear convergence rates as a function of the stepsize, momentum parameter and the noise variance; recovering the accelerated rates in the noiseless case and quantifying the level of noise that can be tolerated to achieve a given performance. To the best of our knowledge, these are the first linear convergence results for stochastic momentum methods under the stochastic oracle model. We also develop finer results for the special case of quadratic objectives, extend our results to the APG method and weakly convex functions showing accelerated rates when the noise magnitude is sufficiently small.
****
Active Embedding Search via Noisy Paired Comparisons
Gregory Canal, Andy Massimino, Mark Davenport, Christopher Rozell
Suppose that we wish to estimate a user's preference vector w from paired comparisons of the form “does user w prefer item p or item q?,” where both the user and items are embedded in a low-dimensional Euclidean space with distances that reflect user and item similarities. Such observations arise in numerous settings, including psychometrics and psychology experiments, search tasks, advertising, and recommender systems. In such tasks, queries can be extremely costly and subject to varying levels of response noise; thus, we aim to actively choose pairs that are most informative given the results of previous comparisons. We provide new theoretical insights into the benefits and challenges of greedy information maximization in this setting, and develop two novel strategies that maximize lower bounds on information gain and are simpler to analyze and compute respectively. We use simulated responses from a real-world dataset to validate our strategies through their similar performance to greedy information maximization, and their superior preference estimation over state-of-the-art selection methods as well as random queries.
****
Dynamic Learning with Frequent New Product Launches: A Sequential Multinomial Logit Bandit Problem
Junyu Cao, Wei Sun
Motivated by the phenomenon that companies introduce new products to keep abreast with customers' rapidly changing tastes, we consider a novel online learning setting where a profit-maximizing seller needs to learn customers' preferences through offering recommendations, which may contain existing products and new products that are launched in the middle of a selling period. We propose a sequential multinomial logit (SMNL) model to characterize customers' behavior when product recommendations are presented in tiers. For the offline version with known customers' preferences, we propose a polynomial-time algorithm and characterize the properties of the optimal tiered product recommendation. For the online problem, we propose a learning algorithm and quantify its regret bound. Moreover, we extend the setting to incorporate a constraint which ensures every new product is learned to a given accuracy. Our results demonstrate the tier structure can be used to mitigate the risks associated with learning new products.

****

Competing Against Nash Equilibria in Adversarially Changing Zero-Sum Games

Adrian Rivera Cardoso, Jacob Abernethy, He Wang, Huan Xu

We study the problem of repeated play in a zero-sum game in which the payoff matrix may change, in a possibly adversarial fashion, on each round; we call these Online Matrix Games. Finding the Nash Equilibrium (NE) of a two player zero-sum game is core to many problems in statistics, optimization, and economics, and for a fixed game matrix this can be easily reduced to solving a linear program. But when the payoff matrix evolves over time our goal is to find a sequential algorithm that can compete with, in a certain sense, the NE of the long-term-averaged payoff matrix. We design an algorithm with small NE regret–that is, we ensure that the long-term payoff of both players is close to minimax optimum in hindsight. Our algorithm achieves near-optimal dependence with respect to the number of rounds and depends poly-logarithmically on the number of available actions of the players. Additionally, we show that the naive reduction, where each player simply minimizes its own regret, fails to achieve the stated objective regardless of which algorithm is used. Lastly, we consider the so-called bandit setting, where the feedback is significantly limited, and we provide an algorithm with small NE regret using one-point estimates of each payoff matrix.

****

Automated Model Selection with Bayesian Quadrature

Henry Chai, Jean-Francois Ton, Michael A. Osborne, Roman Garnett

We present a novel technique for tailoring Bayesian quadrature (BQ) to model selection. The state-of-the-art for comparing the evidence of multiple models relies on Monte Carlo methods, which converge slowly and are unreliable for computationally expensive models. Although previous research has shown that BQ offers sample efficiency superior to Monte Carlo in computing the evidence of an individual model, applying BQ directly to model comparison may waste computation producing an overly-accurate estimate for the evidence of a clearly poor model. We propose an automated and efficient algorithm for computing the most-relevant quantity for model selection: the posterior model probability. Our technique maximizes the mutual information between this quantity and observations of the models' likelihoods, yielding efficient sample acquisition across disparate model spaces when likelihood observations are limited. Our method produces more-accurate posterior estimates using fewer likelihood evaluations than standard Bayesian quadrature and Monte Carlo estimators, as we demonstrate on synthetic and real-world examples.

****

Learning Action Representations for Reinforcement Learning

Yash Chandak, Georgios Theocharous, James Kostas, Scott Jordan, Philip Thomas

Most model-free reinforcement learning methods leverage state representations (embeddings) for generalization, but either ignore structure in the space of actions or assume the structure is provided a priori. We show how a policy can be decomposed into a component that acts in a low-dimensional space of action representations and a component that transforms these representations into actual actions. These representations improve generalization over large, finite action sets by allowing the agent to infer the outcomes of actions similar to actions already taken. We provide an algorithm to both learn and use action representations and provide conditions for its convergence. The efficacy of the proposed method is demonstrated on large-scale real-world problems.

****

Dynamic Measurement Scheduling for Event Forecasting using Deep RL

Chun-Hao Chang, Mingjie Mai, Anna Goldenberg

Imagine a patient in critical condition. What and when should be measured to forecast detrimental events, especially under the budget constraints? We answer this question by deep reinforcement learning (RL) that jointly minimizes the measurement cost and maximizes predictive gain, by scheduling strategically-timed measurements. We learn our policy to be dynamically dependent on the patient's health history. To scale our framework to exponentially large action space, we distribute our reward in a sequential setting that makes the learning easier. In our s

imulation, our policy outperforms heuristic-based scheduling with higher predictive gain and lower cost. In a real-world ICU mortality prediction task (MIMIC3), our policies reduce the total number of measurements by 31% or improve predictive gain by a factor of 3 as compared to physicians, under the off-policy policy evaluation.
****
On Symmetric Losses for Learning from Corrupted Labels
Nontawat Charoenphakdee, Jongyeong Lee, Masashi Sugiyama
This paper aims to provide a better understanding of a symmetric loss. First, we emphasize that using a symmetric loss is advantageous in the balanced error rate (BER) minimization and area under the receiver operating characteristic curve (AUC) maximization from corrupted labels. Second, we prove general theoretical properties of symmetric losses, including a classification-calibration condition, excess risk bound, conditional risk minimizer, and AUC-consistency condition. Third, since all nonnegative symmetric losses are non-convex, we propose a convex barrier hinge loss that benefits significantly from the symmetric condition, although it is not symmetric everywhere. Finally, we conduct experiments to validate the relevance of the symmetric condition.
****
Online learning with kernel losses
Niladri Chatterji, Aldo Pacchiano, Peter Bartlett
We present a generalization of the adversarial linear bandits framework, where the underlying losses are kernel functions (with an associated reproducing kernel Hilbert space) rather than linear functions. We study a version of the exponential weights algorithm and bound its regret in this setting. Under conditions on the eigen-decay of the kernel we provide a sharp characterization of the regret for this algorithm. When we have polynomial eigen-decay ($\mu_j \le \mathcal{O}(j^{-\beta})$), we find that the regret is bounded by $\mathcal{R}_n \le \mathcal{O}(n^{\beta/2(\beta-1)})$. While under the assumption of exponential eigen-decay ($\mu_j \le \mathcal{O}(e^{-\beta j })$) we get an even tighter bound on the regret $\mathcal{R}_n \le \tilde{\mathcal{O}}(n^{1/2})$. When the eigen-decay is polynomial we also show a non-matching minimax lower bound on the regret of $\mathcal{R}_n \ge \Omega(n^{(\beta+1)/2\beta})$ and a lower bound of $\mathcal{R}_n \ge \Omega(n^{1/2})$ when the decay in the eigen-values is exponentially fast. We also study the full information setting when the underlying losses are kernel functions and present an adapted exponential weights algorithm and a conditional gradient descent algorithm.
****
Neural Network Attributions: A Causal Perspective
Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, Vineeth N Balasubramanian
We propose a new attribution method for neural networks developed using ■rst principles of causality (to the best of our knowledge, the ■rst such). The neural network architecture is viewed as a Structural Causal Model, and a methodology to compute the causal effect of each feature on the output is presented. With reasonable assumptions on the causal structure of the input data, we propose algorithms to ef■ciently compute the causal effects, as well as scale the approach to data with large dimensionality. We also show how this method can be used for recurrent neural networks. We report experimental results on both simulated and real datasets showcasing the promise and usefulness of the proposed algorithm.
****
PAC Identification of Many Good Arms in Stochastic Multi-Armed Bandits
Arghya Roy Chaudhuri, Shivaram Kalyanakrishnan
We consider the problem of identifying any k out of the best m arms in an n-armed stochastic multi-armed bandit; framed in the PAC setting, this particular problem generalises both the problem of “best subset selection” (Kalyanakrishnan & Stone, 2010) and that of selecting “one out of the best m” arms (Roy Chaudhuri & Kalyanakrishnan, 2017). We present a lower bound on the worst-case sample complexity for general k, and a fully sequential PAC algorithm, LUCB-k-m, which is more sample-efficient on easy instances. Also, extending our analysis to infinite-a

rmed bandits, we present a PAC algorithm that is independent of n, which identifies an arm from the best ρ fraction of arms using at most an additive poly-log number of samples than compared to the lower bound, thereby improving over Roy Chaudhuri & Kalyanakrishnan (2017) and Aziz et al. (2018). The problem of identifying k > 1 distinct arms from the best ρ fraction is not always well-defined; for a special class of this problem, we present lower and upper bounds. Finally, through a reduction, we establish a relation between upper bounds for the “one out of the best ρ” problem for infinite instances and the “one out of the best m” problem for finite instances. We conjecture that it is more efficient to solve “small” finite instances using the latter formulation, rather than going through the former.
****
Nearest Neighbor and Kernel Survival Analysis: Nonasymptotic Error Bounds and Strong Consistency Rates
George Chen
We establish the first nonasymptotic error bounds for Kaplan-Meier-based nearest neighbor and kernel survival probability estimators where feature vectors reside in metric spaces. Our bounds imply rates of strong consistency for these nonparametric estimators and, up to a log factor, match an existing lower bound for conditional CDF estimation. Our proof strategy also yields nonasymptotic guarantees for nearest neighbor and kernel variants of the Nelson-Aalen cumulative hazards estimator. We experimentally compare these methods on four datasets. We find that for the kernel survival estimator, a good choice of kernel is one learned using random survival forests.
****
Stein Point Markov Chain Monte Carlo
Wilson Ye Chen, Alessandro Barp, Francois-Xavier Briol, Jackson Gorham, Mark Girolami, Lester Mackey, Chris Oates
An important task in machine learning and statistics is the approximation of a probability measure by an empirical measure supported on a discrete point set. Stein Points are a class of algorithms for this task, which proceed by sequentially minimising a Stein discrepancy between the empirical measure and the target and, hence, require the solution of a non-convex optimisation problem to obtain each new point. This paper removes the need to solve this optimisation problem by, instead, selecting each new point based on a Markov chain sample path. This significantly reduces the computational cost of Stein Points and leads to a suite of algorithms that are straightforward to implement. The new algorithms are illustrated on a set of challenging Bayesian inference problems, and rigorous theoretical guarantees of consistency are established.
****
Particle Flow Bayes’ Rule
Xinshi Chen, Hanjun Dai, Le Song
We present a particle flow realization of Bayes’ rule, where an ODE-based neural operator is used to transport particles from a prior to its posterior after a new observation. We prove that such an ODE operator exists. Its neural parameterization can be trained in a meta-learning framework, allowing this operator to reason about the effect of an individual observation on the posterior, and thus generalize across different priors, observations and to sequential Bayesian inference. We demonstrated the generalization ability of our particle flow Bayes operator in several canonical and high dimensional examples.
****
Proportionally Fair Clustering
Xingyu Chen, Brandon Fain, Liang Lyu, Kamesh Munagala
We extend the fair machine learning literature by considering the problem of proportional centroid clustering in a metric context. For clustering n points with k centers, we define fairness as proportionality to mean that any n/k points are entitled to form their own cluster if there is another center that is closer in distance for all n/k points. We seek clustering solutions to which there are no such justified complaints from any subsets of agents, without assuming any a priori notion of protected subsets. We present and analyze algorithms to efficient

ly compute, optimize, and audit proportional solutions. We conclude with an empirical examination of the tradeoff between proportional solutions and the k-means objective.
****
Information-Theoretic Considerations in Batch Reinforcement Learning
Jinglin Chen, Nan Jiang
Value-function approximation methods that operate in batch mode have foundational importance to reinforcement learning (RL). Finite sample guarantees for these methods often crucially rely on two types of assumptions: (1) mild distribution shift, and (2) representation conditions that are stronger than realizability. However, the necessity (“why do we need them?”) and the naturalness (“when do they hold?”) of such assumptions have largely eluded the literature. In this paper, we revisit these assumptions and provide theoretical results towards answering the above questions, and make steps towards a deeper understanding of value-function approximation.
****
Generative Adversarial User Model for Reinforcement Learning Based Recommendation System
Xinshi Chen, Shuang Li, Hui Li, Shaohua Jiang, Yuan Qi, Le Song
There are great interests as well as many challenges in applying reinforcement learning (RL) to recommendation systems. In this setting, an online user is the environment; neither the reward function nor the environment dynamics are clearly defined, making the application of RL challenging. In this paper, we propose a novel model-based reinforcement learning framework for recommendation systems, where we develop a generative adversarial network to imitate user behavior dynamics and learn her reward function. Using this user model as the simulation environment, we develop a novel Cascading DQN algorithm to obtain a combinatorial recommendation policy which can handle a large number of candidate items efficiently. In our experiments with real data, we show this generative adversarial user model can better explain user behavior than alternatives, and the RL policy based on this model can lead to a better long-term reward for the user and higher click rate for the system.
****
Understanding and Utilizing Deep Neural Networks Trained with Noisy Labels
Pengfei Chen, Ben Ben Liao, Guangyong Chen, Shengyu Zhang
Noisy labels are ubiquitous in real-world datasets, which poses a challenge for robustly training deep neural networks (DNNs) as DNNs usually have the high capacity to memorize the noisy labels. In this paper, we find that the test accuracy can be quantitatively characterized in terms of the noise ratio in datasets. In particular, the test accuracy is a quadratic function of the noise ratio in the case of symmetric noise, which explains the experimental findings previously published. Based on our analysis, we apply cross-validation to randomly split noisy datasets, which identifies most samples that have correct labels. Then we adopt the Co-teaching strategy which takes full advantage of the identified samples to train DNNs robustly against noisy labels. Compared with extensive state-of-the-art methods, our strategy consistently improves the generalization performance of DNNs under both synthetic and real-world training noise.
****
A Gradual, Semi-Discrete Approach to Generative Network Training via Explicit Wasserstein Minimization
Yucheng Chen, Matus Telgarsky, Chao Zhang, Bolton Bailey, Daniel Hsu, Jian Peng
This paper provides a simple procedure to fit generative networks to target distributions, with the goal of a small Wasserstein distance (or other optimal transport costs). The approach is based on two principles: (a) if the source randomness of the network is a continuous distribution (the "semi-discrete" setting), then the Wasserstein distance is realized by a deterministic optimal transport mapping; (b) given an optimal transport mapping between a generator network and a target distribution, the Wasserstein distance may be decreased via a regression between the generated data and the mapped target points. The procedure here therefore alternates these two steps, forming an optimal transport and regressing aga

inst it, gradually adjusting the generator network towards the target distributi on. Mathematically, this approach is shown to minimize the Wasserstein distance to both the empirical target distribution, and also its underlying population co unterpart. Empirically, good performance is demonstrated on the training and tes ting sets of the MNIST and Thin-8 data. The paper closes with a discussion of th e unsuitability of the Wasserstein distance for certain tasks, as has been ident ified in prior work (Arora et al., 2017; Huang et al., 2017).
****
Transferability vs. Discriminability: Batch Spectral Penalization for Adversaria l Domain Adaptation
Xinyang Chen, Sinan Wang, Mingsheng Long, Jianmin Wang
Adversarial domain adaptation has made remarkable advances in learning transfera ble representations for knowledge transfer across domains. While adversarial lea rning strengthens the feature transferability which the community focuses on, it s impact on the feature discriminability has not been fully explored. In this pa per, a series of experiments based on spectral analysis of the feature represent ations have been conducted, revealing an unexpected deterioration of the discrim inability while learning transferable features adversarially. Our key finding is that the eigenvectors with the largest singular values will dominate the featur e transferability. As a consequence, the transferability is enhanced at the expe nse of over penalization of other eigenvectors that embody rich structures cruci al for discriminability. Towards this problem, we present Batch Spectral Penaliz ation (BSP), a general approach to penalizing the largest singular values so tha t other eigenvectors can be relatively strengthened to boost the feature discrim inability. Experiments show that the approach significantly improves upon repres entative adversarial domain adaptation methods to yield state of the art results .
****
Fast Incremental von Neumann Graph Entropy Computation: Theory, Algorithm, and A pplications
Pin-Yu Chen, Lingfei Wu, Sijia Liu, Indika Rajapakse
The von Neumann graph entropy (VNGE) facilitates measurement of information dive rgence and distance between graphs in a graph sequence. It has been successfully applied to various learning tasks driven by network-based data. While effective , VNGE is computationally demanding as it requires the full eigenspectrum of the graph Laplacian matrix. In this paper, we propose a new computational framework , Fast Incremental von Neumann Graph EntRopy (FINGER), which approaches VNGE wit h a performance guarantee. FINGER reduces the cubic complexity of VNGE to linear complexity in the number of nodes and edges, and thus enables online computatio n based on incremental graph changes. We also show asymptotic equivalence of FIN GER to the exact VNGE, and derive its approximation error bounds. Based on FINGE R, we propose efficient algorithms for computing Jensen-Shannon distance between graphs. Our experimental results on different random graph models demonstrate t he computational efficiency and the asymptotic equivalence of FINGER. In additio n, we apply FINGER to two real-world applications and one synthesized anomaly de tection dataset, and corroborate its superior performance over seven baseline gr aph similarity methods.
****
Katalyst: Boosting Convex Katayusha for Non-Convex Problems with a Large Conditi on Number
Zaiyi Chen, Yi Xu, Haoyuan Hu, Tianbao Yang
An important class of non-convex objectives that has wide applications in machin e learning consists of a sum of n smooth functions and a non-smooth convex fun ction. Tremendous studies have been devoted to conquering these problems by leve raging one of the two types of variance reduction techniques, i.e., SVRG-type th at computes a full gradient occasionally and SAGA-type that maintains n stocha stic gradients at every iteration. In practice, SVRG-type is preferred to SAGA-t ype due to its potentially less memory costs. An interesting question that has b een largely ignored is how to improve the complexity of variance reduction metho ds for problems with a large condition number that measures the degree to which

the objective is close to a convex function. In this paper, we present a simple but non-trivial boosting of a state-of-the-art SVRG-type method for convex problems (namely Katyusha) to enjoy an improved complexity for solving non-convex problems with a large condition number (that is close to a convex function). To the best of our knowledge, its complexity has the best dependence on n and the degree of non-convexity, and also matches that of a recent SAGA-type accelerated stochastic algorithm for a constrained non-convex smooth optimization problem.
****
Multivariate-Information Adversarial Ensemble for Scalable Joint Distribution Matching
Ziliang Chen, Zhanfu Yang, Xiaoxi Wang, Xiaodan Liang, Xiaopeng Yan, Guanbin Li, Liang Lin
A broad range of cross-m-domain generation researches boil down to matching a joint distribution by deep generative models (DGMs). Hitherto algorithms excel in pairwise domains while as m increases, remain struggling to scale themselves to ■t a joint distribution. In this paper, we propose a domain-scalable DGM, i.e., MMI-ALI for m-domain joint distribution matching. As an m-domain ensemble model of ALIs (Dumoulin et al., 2016), MMI-ALI is adversarially trained with maximizing Multivariate Mutual Information (MMI) w.r.t. joint variables of each pair of domains and their shared feature. The negative MMIs are upper bounded by a series of feasible losses provably leading to matching m-domain joint distributions. MMI-ALI linearly scales as m increases and thus, strikes a right balance between ef■cacy and scalability. We evaluate MMI-ALI in diverse challenging m-domain scenarios and verify its superiority.
****
Robust Decision Trees Against Adversarial Examples
Hongge Chen, Huan Zhang, Duane Boning, Cho-Jui Hsieh
Although adversarial examples and model robust-ness have been extensively studied in the context of neural networks, research on this issue in tree-based models and how to make tree-based models robust against adversarial examples is still limited. In this paper, we show that tree-based models are also vulnerable to adversarial examples and develop a novel algorithm to learn robust trees. At its core, our method aims to optimize the performance under the worst-case perturbation of input features, which leads to a max-min saddle point problem. Incorporating this saddle point objective into the decision tree building procedure is non-trivial due to the discrete nature of trees{—}a naive approach to finding the best split according to this saddle point objective will take exponential time. To make our approach practical and scalable, we propose efficient tree building algorithms by approximating the inner minimizer in the saddlepoint problem, and present efficient implementations for classical information gain based trees as well as state-of-the-art tree boosting systems such as XGBoost. Experimental results on real world datasets demonstrate that the proposed algorithms can significantly improve the robustness of tree-based models against adversarial examples.
****
RaFM: Rank-Aware Factorization Machines
Xiaoshuang Chen, Yin Zheng, Jiaxing Wang, Wenye Ma, Junzhou Huang
Fatorization machines (FM) are a popular model class to learn pairwise interactions by a low-rank approximation. Different from existing FM-based approaches which use a fixed rank for all features, this paper proposes a Rank-Aware FM (RaFM) model which adopts pairwise interactions from embeddings with different ranks. The proposed model achieves a better performance on real-world datasets where different features have significantly varying frequencies of occurrences. Moreover, we prove that the RaFM model can be stored, evaluated, and trained as efficiently as one single FM, and under some reasonable conditions it can be even significantly more efficient than FM. RaFM improves the performance of FMs in both regression tasks and classification tasks while incurring less computational burden, therefore also has attractive potential in industrial applications.
****
Control Regularization for Reduced Variance Reinforcement Learning
Richard Cheng, Abhinav Verma, Gabor Orosz, Swarat Chaudhuri, Yisong Yue, Joel Bu

rdick
Dealing with high variance is a significant challenge in model-free reinforcement learning (RL). Existing methods are unreliable, exhibiting high variance in performance from run to run using different initializations/seeds. Focusing on problems arising in continuous control, we propose a functional regularization approach to augmenting model-free RL. In particular, we regularize the behavior of the deep policy to be similar to a policy prior, i.e., we regularize in function space. We show that functional regularization yields a bias-variance trade-off, and propose an adaptive tuning strategy to optimize this trade-off. When the policy prior has control-theoretic stability guarantees, we further show that this regularization approximately preserves those stability guarantees throughout learning. We validate our approach empirically on a range of settings, and demonstrate significantly reduced variance, guaranteed dynamic stability, and more efficient learning than deep RL alone.
****

Predictor-Corrector Policy Optimization
Ching-An Cheng, Xinyan Yan, Nathan Ratliff, Byron Boots
We present a predictor-corrector framework, called PicCoLO, that can transform a first-order model-free reinforcement or imitation learning algorithm into a new hybrid method that leverages predictive models to accelerate policy learning. The new “PicCoLOed” algorithm optimizes a policy by recursively repeating two steps: In the Prediction Step, the learner uses a model to predict the unseen future gradient and then applies the predicted estimate to update the policy; in the Correction Step, the learner runs the updated policy in the environment, receives the true gradient, and then corrects the policy using the gradient error. Unlike previous algorithms, PicCoLO corrects for the mistakes of using imperfect predicted gradients and hence does not suffer from model bias. The development of PicCoLO is made possible by a novel reduction from predictable online learning to adversarial online learning, which provides a systematic way to modify existing first-order algorithms to achieve the optimal regret with respect to predictable information. We show, in both theory and simulation, that the convergence rate of several first-order model-free algorithms can be improved by PicCoLO.
****

Variational Inference for sparse network reconstruction from count data
Julien Chiquet, Stephane Robin, Mahendra Mariadassou
Networks provide a natural yet statistically grounded way to depict and understand how a set of entities interact. However, in many situations interactions are not directly observed and the network needs to be reconstructed based on observations collected for each entity. Our work focuses on the situation where these observations consist of counts. A typical example is the reconstruction of an ecological network based on abundance data. In this setting, the abundance of a set of species is collected in a series of samples and/or environments and we aim at inferring direct interactions between the species. The abundances at hand can be, for example, direct counts of individuals (ecology of macro-organisms) or read counts resulting from metagenomic sequencing (microbial ecology). Whatever the approach chosen to infer such a network, it has to account for the peculiaraties of the data at hand. The first, obvious one, is that the data are counts, i.e. non continuous. Also, the observed counts often vary over many orders of magnitude and are more dispersed than expected under a simple model, such as the Poisson distribution. The observed counts may also result from different sampling efforts in each sample and/or for each entity, which hampers direct comparison. Furthermore, because the network is supposed to reveal only direct interactions, it is highly desirable to account for covariates describing the environment to avoid spurious edges. Many methods of network reconstruction from count data have been proposed. In the context of microbial ecology, most methods (SparCC, REBACCA, SPIEC-EASI, gCODA, BanOCC) rely on a two-step strategy: transform the counts to pseudo Gaussian observations using simple transforms before moving back to the setting of Gaussian Graphical Models, for which state of the art methods exist to infer the network, but only in a Gaussian world. In this work, we consider instead a full-fledged probabilistic model with a latent layer where the counts f

ollow Poisson distributions, conditional to latent (hidden) Gaussian correlated variables. In this model, known as Poisson log-normal (PLN), the dependency structure is completely captured by the latent layer and we model counts, rather than transformations thereof. To our knowledge, the PLN framework is quite new and has only been used by two other recent methods (Mint and plnDAG) to reconstruct networks from count data. In this work, we use the same mathematical framework but adopt a different optimization strategy which alleviates the whole optimization process. We also fully exploit the connection between the PLN framework and generalized linear models to account for the peculiarities of microbiological data sets. The network inference step is done as usual by adding sparsity inducing constraints on the inverse covariance matrix of the latent Gaussian vector to select only the most important interactions between species. Unlike the usual Gaussian setting, the penalized likelihood is generally not tractable in this framework. We resort instead to a variational approximation for parameter inference and solve the corresponding optimization problem by alternating a gradient descent on the variational parameters and a graphical-Lasso step on the covariance matrix. We also select the sparsity parameter using the resampling-based StARS procedure. We show that the sparse PLN approach has better performance than existing methods on simulated datasets and that it extracts relevant signal from microbial ecology datasets. We also show that the inference scales to datasets made up of hundred of species and samples, in line with other methods in the field. In short, our contributions to the field are the following: we extend the use of PLN distributions in network inference by (i) accounting for covariates and offset and thus removing some spurious edges induced by confounding factors, (ii) accounting for different sampling effort to integrate data sets from different sources and thus infer interactions between different types of organisms (e.g. bacteria - fungi), (iii) developing an inference procedure based on the iterative optimization of a well defined objective function. Our objective function is a provable lower bound of the observed likelihood and our procedure accounts for the uncertainty associated with the estimation of the latent variable, unlike the algorithm presented in Mint and plnDAG.
****

Random Walks on Hypergraphs with Edge-Dependent Vertex Weights

Uthsav Chitra, Benjamin Raphael

Hypergraphs are used in machine learning to model higher-order relationships in data. While spectral methods for graphs are well-established, spectral theory for hypergraphs remains an active area of research. In this paper, we use random walks to develop a spectral theory for hypergraphs with edge-dependent vertex weights: hypergraphs where every vertex v has a weight $\gamma_e(v)$ for each incident hyperedge e that describes the contribution of v to the hyperedge e. We derive a random walk-based hypergraph Laplacian, and bound the mixing time of random walks on such hypergraphs. Moreover, we give conditions under which random walks on such hypergraphs are equivalent to random walks on graphs. As a corollary, we show that current machine learning methods that rely on Laplacians derived from random walks on hypergraphs with edge-independent vertex weights do not utilize higher-order relationships in the data. Finally, we demonstrate the advantages of hypergraphs with edge-dependent vertex weights on ranking applications using real-world datasets.
****

Neural Joint Source-Channel Coding

Kristy Choi, Kedar Tatwawadi, Aditya Grover, Tsachy Weissman, Stefano Ermon

For reliable transmission across a noisy communication channel, classical results from information theory show that it is asymptotically optimal to separate out the source and channel coding processes. However, this decomposition can fall short in the finite bit-length regime, as it requires non-trivial tuning of hand-crafted codes and assumes infinite computational power for decoding. In this work, we propose to jointly learn the encoding and decoding processes using a new discrete variational autoencoder model. By adding noise into the latent codes to simulate the channel during training, we learn to both compress and error-correct given a fixed bit-length and computational budget. We obtain codes that are no

t only competitive against several separation schemes, but also learn useful rob
ust representations of the data for downstream tasks such as classification. Fin
ally, inference amortization yields an extremely fast neural decoder, almost an
order of magnitude faster compared to standard decoding methods based on iterati
ve belief propagation.
****
Beyond Backprop: Online Alternating Minimization with Auxiliary Variables
Anna Choromanska, Benjamin Cowen, Sadhana Kumaravel, Ronny Luss, Mattia Rigotti,
 Irina Rish, Paolo Diachille, Viatcheslav Gurev, Brian Kingsbury, Ravi Tejwani,
Djallel Bouneffouf
Despite significant recent advances in deep neural networks, training them remai
ns a challenge due to the highly non-convex nature of the objective function. St
ate-of-the-art methods rely on error backpropagation, which suffers from several
 well-known issues, such as vanishing and exploding gradients, inability to hand
le non-differentiable nonlinearities and to parallelize weight-updates across la
yers, and biological implausibility. These limitations continue to motivate expl
oration of alternative training algorithms, including several recently proposed
auxiliary-variable methods which break the complex nested objective function int
o local subproblems. However, those techniques are mainly offline (batch), which
 limits their applicability to extremely large datasets, as well as to online, c
ontinual or reinforcement learning. The main contribution of our work is a novel
 online (stochastic/mini-batch) alternating minimization (AM) approach for train
ing deep neural networks, together with the first theoretical convergence guaran
tees for AM in stochastic settings and promising empirical results on a variety
of architectures and datasets.
****
Unifying Orthogonal Monte Carlo Methods
Krzysztof Choromanski, Mark Rowland, Wenyu Chen, Adrian Weller
Many machine learning methods making use of Monte Carlo sampling in vector space
s have been shown to be improved by conditioning samples to be mutually orthogon
al. Exact orthogonal coupling of samples is computationally intensive, hence app
roximate methods have been of great interest. In this paper, we present a unifyi
ng perspective of many approximate methods by considering Givens transformations
, propose new approximate methods based on this framework, and demonstrate the ■
rst statistical guarantees for families of approximate methods in kernel approxi
mation. We provide extensive empirical evaluations with guidance for practitione
rs.
****
Probability Functional Descent: A Unifying Perspective on GANs, Variational Infe
rence, and Reinforcement Learning
Casey Chu, Jose Blanchet, Peter Glynn
The goal of this paper is to provide a unifying view of a wide range of problems
 of interest in machine learning by framing them as the minimization of function
als defined on the space of probability measures. In particular, we show that ge
nerative adversarial networks, variational inference, and actor-critic methods i
n reinforcement learning can all be seen through the lens of our framework. We t
hen discuss a generic optimization algorithm for our formulation, called probabi
lity functional descent (PFD), and show how this algorithm recovers existing met
hods developed independently in the settings mentioned earlier.
****
MeanSum: A Neural Model for Unsupervised Multi-Document Abstractive Summarizatio
n
Eric Chu, Peter Liu
Abstractive summarization has been studied using neural sequence transduction me
thods with datasets of large, paired document-summary examples. However, such da
tasets are rare and the models trained from them do not generalize to other doma
ins. Recently, some progress has been made in learning sequence-to-sequence mapp
ings with only unpaired examples. In our work, we consider the setting where the
re are only documents (product or business reviews) with no summaries provided,
and propose an end-to-end, neural model architecture to perform unsupervised abs

tractive summarization. Our proposed model consists of an auto-encoder where the mean of the representations of the input reviews decodes to a reasonable summary-review. We consider variants of the proposed architecture and perform an ablation study to show the importance of specific components. We show through metrics and human evaluation that the generated summaries are highly abstractive, fluent, relevant, and representative of the average sentiment of the input reviews. Finally, we collect a ground-truth evaluation dataset and show that our model outperforms a strong extractive baseline.
****

Weak Detection of Signal in the Spiked Wigner Model
Hye Won Chung, Ji Oon Lee
We consider the problem of detecting the presence of the signal in a rank-one signal-plus-noise data matrix. In case the signal-to-noise ratio is under the threshold below which a reliable detection is impossible, we propose a hypothesis test based on the linear spectral statistics of the data matrix. When the noise is Gaussian, the error of the proposed test is optimal as it matches the error of the likelihood ratio test that minimizes the sum of the Type-I and Type-II errors. The test is data-driven and does not depend on the distribution of the signal or the noise. If the density of the noise is known, it can be further improved by an entrywise transformation to lower the error of the test.
****

New results on information theoretic clustering
Ferdinando Cicalese, Eduardo Laber, Lucas Murtinho
We study the problem of optimizing the clustering of a set of vectors when the quality of the clustering is measured by the Entropy or the Gini impurity measure. Our results contribute to the state of the art both in terms of best known approximation guarantees and inapproximability bounds: (i) we give the first polynomial time algorithm for Entropy impurity based clustering with approximation guarantee independent of the number of vectors and (ii) we show that the problem of clustering based on entropy impurity does not admit a PTAS. This also implies an inapproximability result in information theoretic clustering for probability distributions closing a problem left open in [Chaudhury and McGregor, COLT08] and [Ackermann et al., ECCC11]. We also report experiments with a new clustering method that was designed on top of the theoretical tools leading to the above results. These experiments suggest a practical applicability for our method, in particular, when the number of clusters is large.
****

Sensitivity Analysis of Linear Structural Causal Models
Carlos Cinelli, Daniel Kumor, Bryant Chen, Judea Pearl, Elias Bareinboim
Causal inference requires assumptions about the data generating process, many of which are unverifiable from the data. Given that some causal assumptions might be uncertain or disputed, formal methods are needed to quantify how sensitive research conclusions are to violations of those assumptions. Although an extensive literature exists on the topic, most results are limited to specific model structures, while a general-purpose algorithmic framework for sensitivity analysis is still lacking. In this paper, we develop a formal, systematic approach to sensitivity analysis for arbitrary linear Structural Causal Models (SCMs). We start by formalizing sensitivity analysis as a constrained identification problem. We then develop an efficient, graph-based identification algorithm that exploits non-zero constraints on both directed and bidirected edges. This allows researchers to systematically derive sensitivity curves for a target causal quantity with an arbitrary set of path coefficients and error covariances as sensitivity parameters. These results can be used to display the degree to which violations of causal assumptions affect the target quantity of interest, and to judge, on scientific grounds, whether problematic degrees of violations are plausible.
****

Dimensionality Reduction for Tukey Regression
Kenneth Clarkson, Ruosong Wang, David Woodruff
We give the first dimensionality reduction methods for the overconstrained Tukey regression problem. The Tukey loss function $\|y\|_M = \sum_i M(y_i)$ has $M(y_

i) \approx |y_i|^p$ for residual errors y_i smaller than a prescribed threshol
d τ, but $M(y_i)$ becomes constant for errors $|y_i| > \tau$. Our results d
epend on a new structural result, proven constructively, showing that for any $d
$-dimensional subspace $L \subset \mathbb{R}^n$, there is a fixed bounded-size s
ubset of coordinates containing, for every $y \in L$, all the large coordinates,
 with respect to the Tukey loss function, of y. Our methods reduce a given Tuk
ey regression problem to a smaller weighted version, whose solution is a provabl
y good approximate solution to the original problem. Our reductions are fast, si
mple and easy to implement, and we give empirical results demonstrating their pr
acticality, using existing heuristic solvers for the small versions. We also giv
e exponential-time algorithms giving provably good solutions, and hardness resul
ts suggesting that a significant speedup in the worst case is unlikely.
****
On Medians of (Randomized) Pairwise Means
Pierre Laforgue, Stephan Clemencon, Patrice Bertail
Tournament procedures, recently introduced in the literature, offer an appealing
 alternative, from a theoretical perspective at least, to the principle of Empir
ical Risk Minimization in machine learning. Statistical learning by Median-of-Me
ans (MoM) basically consists in segmenting the training data into blocks of equa
l size and comparing the statistical performance of every pair of candidate deci
sion rules on each data block: that with highest performance on the majority of
the blocks is declared as the winner. In the context of nonparametric regression
, functions having won all their duels have been shown to outperform empirical r
isk minimizers w.r.t. the mean squared error under minimal assumptions, while ex
hibiting robustness properties. It is the purpose of this paper to extend this a
pproach, in order to address other learning problems in particular, for which th
e performance criterion takes the form of an expectation over pairs of observati
ons rather than over one single observation, as may be the case in pairwise rank
ing, clustering or metric learning. Precisely, it is proved here that the bounds
 achieved by MoM are essentially conserved when the blocks are built by means of
 independent sampling without replacement schemes instead of a simple segmentati
on. These results are next extended to situations where the risk is related to a
 pairwise loss function and its empirical counterpart is of the form of a U-st
atistic. Beyond theoretical results guaranteeing the performance of the learning
/estimation methods proposed, some numerical experiments provide empirical evide
nce of their relevance in practice.
****
Quantifying Generalization in Reinforcement Learning
Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, John Schulman
In this paper, we investigate the problem of overfitting in deep reinforcement l
earning. Among the most common benchmarks in RL, it is customary to use the same
 environments for both training and testing. This practice offers relatively lit
tle insight into an agent’s ability to generalize. We address this issue by usin
g procedurally generated environments to construct distinct training and test se
ts. Most notably, we introduce a new environment called CoinRun, designed as a b
enchmark for generalization in RL. Using CoinRun, we find that agents overfit to
 surprisingly large training sets. We then show that deeper convolutional archit
ectures improve generalization, as do methods traditionally found in supervised
learning, including L2 regularization, dropout, data augmentation and batch norm
alization.
****
Empirical Analysis of Beam Search Performance Degradation in Neural Sequence Mod
els
Eldan Cohen, Christopher Beck
Beam search is the most popular inference algorithm for decoding neural sequence
 models. Unlike greedy search, beam search allows for non-greedy local decisions
 that can potentially lead to a sequence with a higher overall probability. Howe
ver, work on a number of applications has found that the quality of the highest
probability hypothesis found by beam search degrades with large beam widths. We
perform an empirical study of the behavior of beam search across three sequence

synthesis tasks. We find that increasing the beam width leads to sequences that are disproportionately based on early, very low probability tokens that are followed by a sequence of tokens with higher (conditional) probability. We show that, empirically, such sequences are more likely to have a lower evaluation score than lower probability sequences without this pattern. Using the notion of search discrepancies from heuristic search, we hypothesize that large discrepancies are the cause of the performance degradation. We show that this hypothesis generalizes the previous ones in machine translation and image captioning. To validate our hypothesis, we show that constraining beam search to avoid large discrepancies eliminates the performance degradation.
****
Learning Linear-Quadratic Regulators Efficiently with only $\sqrt{T}$ Regret
Alon Cohen, Tomer Koren, Yishay Mansour
We present the first computationally-efficient algorithm with $\widetilde{O}(\sqrt{T})$ regret for learning in Linear Quadratic Control systems with unknown dynamics. By that, we resolve an open question of Abbasi-Yadkori and Szepesvari (2011) and Dean,Mania, Matni, Recht, and Tu (2018).
****
Certified Adversarial Robustness via Randomized Smoothing
Jeremy Cohen, Elan Rosenfeld, Zico Kolter
We show how to turn any classifier that classifies well under Gaussian noise into a new classifier that is certifiably robust to adversarial perturbations under the L2 norm. While this "randomized smoothing" technique has been proposed before in the literature, we are the first to provide a tight analysis, which establishes a close connection between L2 robustness and Gaussian noise. We use the technique to train an ImageNet classifier with e.g. a certified top-1 accuracy of 49% under adversarial perturbations with L2 norm less than 0.5 (=127/255). Smoothing is the only approach to certifiably robust classification which has been shown feasible on full-resolution ImageNet. On smaller-scale datasets where competing approaches to certified L2 robustness are viable, smoothing delivers higher certified accuracies. The empirical success of the approach suggests that provable methods based on randomization at prediction time are a promising direction for future research into adversarially robust classification.
****
Gauge Equivariant Convolutional Networks and the Icosahedral CNN
Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, Max Welling
The principle of equivariance to symmetry transformations enables a theoretically grounded approach to neural network architecture design. Equivariant networks have shown excellent performance and data efficiency on vision and medical imaging problems that exhibit symmetries. Here we show how this principle can be extended beyond global symmetries to local gauge transformations. This enables the development of a very general class of convolutional neural networks on manifolds that depend only on the intrinsic geometry, and which includes many popular methods from equivariant and geometric deep learning. We implement gauge equivariant CNNs for signals defined on the surface of the icosahedron, which provides a reasonable approximation of the sphere. By choosing to work with this very regular manifold, we are able to implement the gauge equivariant convolution using a single conv2d call, making it a highly scalable and practical alternative to Spherical CNNs. Using this method, we demonstrate substantial improvements over previous methods on the task of segmenting omnidirectional images and global climate patterns.
****
CURIOUS: Intrinsically Motivated Modular Multi-Goal Reinforcement Learning
Cédric Colas, Pierre Fournier, Mohamed Chetouani, Olivier Sigaud, Pierre-Yves Oudeyer
In open-ended environments, autonomous learning agents must set their own goals and build their own curriculum through an intrinsically motivated exploration. They may consider a large diversity of goals, aiming to discover what is controllable in their environments, and what is not. Because some goals might prove easy and some impossible, agents must actively select which goal to practice at any

moment, to maximize their overall mastery on the set of learnable goals. This paper proposes CURIOUS , an algorithm that leverages 1) a modular Universal Value Function Approximator with hindsight learning to achieve a diversity of goals of different kinds within a unique policy and 2) an automated curriculum learning mechanism that biases the attention of the agent towards goals maximizing the absolute learning progress. Agents focus sequentially on goals of increasing complexity, and focus back on goals that are being forgotten. Experiments conducted in a new modular-goal robotic environment show the resulting developmental self-organization of a learning curriculum, and demonstrate properties of robustness to distracting goals, forgetting and changes in body properties.
****
A fully differentiable beam search decoder
Ronan Collobert, Awni Hannun, Gabriel Synnaeve
We introduce a new beam search decoder that is fully differentiable, making it possible to optimize at training time through the inference procedure. Our decoder allows us to combine models which operate at different granularities (e.g. acoustic and language models). It can be used when target sequences are not aligned to input sequences by considering all possible alignments between the two. We demonstrate our approach scales by applying it to speech recognition, jointly training acoustic and word-level language models. The system is end-to-end, with gradients flowing through the whole architecture from the word-level transcriptions. Recent research efforts have shown that deep neural networks with attention-based mechanisms can successfully train an acoustic model from the final transcription, while implicitly learning a language model. Instead, we show that it is possible to discriminatively train an acoustic model jointly with an explicit and possibly pre-trained language model.
****
Scalable Metropolis-Hastings for Exact Bayesian Inference with Large Datasets
Rob Cornish, Paul Vanetti, Alexandre Bouchard-Cote, George Deligiannidis, Arnaud Doucet
Bayesian inference via standard Markov Chain Monte Carlo (MCMC) methods such as Metropolis-Hastings is too computationally intensive to handle large datasets, since the cost per step usually scales like $O(n)$ in the number of data points n. We propose the Scalable Metropolis-Hastings (SMH) kernel that only requires processing on average $O(1)$ or even $O(1/\sqrt{n})$ data points per step. This scheme is based on a combination of factorized acceptance probabilities, procedures for fast simulation of Bernoulli processes, and control variate ideas. Contrary to many MCMC subsampling schemes such as fixed step-size Stochastic Gradient Langevin Dynamics, our approach is exact insofar as the invariant distribution is the true posterior and not an approximation to it. We characterise the performance of our algorithm theoretically, and give realistic and verifiable conditions under which it is geometrically ergodic. This theory is borne out by empirical results that demonstrate overall performance benefits over standard Metropolis-Hastings and various subsampling algorithms.
****
Adjustment Criteria for Generalizing Experimental Findings
Juan Correa, Jin Tian, Elias Bareinboim
Generalizing causal effects from a controlled experiment to settings beyond the particular study population is arguably one of the central tasks found in empirical circles. While a proper design and careful execution of the experiment would support, under mild conditions, the validity of inferences about the population in which the experiment was conducted, two challenges make the extrapolation step to different populations somewhat involved, namely, transportability and sampling selection bias. The former is concerned with disparities in the distributions and causal mechanisms between the domain (i.e., settings, population, environment) where the experiment is conducted and where the inferences are intended; the latter with distortions in the sample’s proportions due to preferential selection of units into the study. In this paper, we investigate the assumptions and machinery necessary for using covariate adjustment to correct for the biases generated by both of these problems, and generalize experimental data to infer caus

al effects in a new domain. We derive complete graphical conditions to determine if a set of covariates is admissible for adjustment in this new setting. Building on the graphical characterization, we develop an efficient algorithm that enumerates all possible admissible sets with poly-time delay guarantee; this can be useful for when some variables are preferred over the others due to different costs or amenability to measurement.
****
Online Learning with Sleeping Experts and Feedback Graphs
Corinna Cortes, Giulia Desalvo, Claudio Gentile, Mehryar Mohri, Scott Yang
We consider the scenario of online learning with sleeping experts, where not all experts are available at each round, and analyze the general framework of learning with feedback graphs, where the loss observations associated with each expert are characterized by a graph. A critical assumption in this framework is that the loss observations and the set of sleeping experts at each round are independent. We first extend the classical sleeping experts algorithm of Kleinberg et al. 2008 to the feedback graphs scenario, and prove matching upper and lower bounds for the sleeping regret of the resulting algorithm under the independence assumption. Our main contribution is then to relax this assumption, present a more general notion of sleeping regret, and derive a general algorithm with strong theoretical guarantees. We apply this new framework to the important scenario of online learning with abstention, where a learner can elect to abstain from making a prediction at the price of a certain cost. We empirically validate our algorithm against multiple online abstention algorithms on several real-world datasets, showing substantial performance improvements.
****
Active Learning with Disagreement Graphs
Corinna Cortes, Giulia Desalvo, Mehryar Mohri, Ningshan Zhang, Claudio Gentile
We present two novel enhancements of an online importance-weighted active learning algorithm IWAL, using the properties of disagreements among hypotheses. The first enhancement, IWALD, prunes the hypothesis set with a more aggressive strategy based on the disagreement graph. We show that IWAL-D improves the generalization performance and the label complexity of the original IWAL, and quantify the improvement in terms of the disagreement graph coefficient. The second enhancement, IZOOM, further improves IWAL-D by adaptively zooming into the current version space and thus reducing the best-in-class error. We show that IZOOM admits favorable theoretical guarantees with the changing hypothesis set. We report experimental results on multiple datasets and demonstrate that the proposed algorithms achieve better test performances than IWAL given the same amount of labeling budget.
****
Shape Constraints for Set Functions
Andrew Cotter, Maya Gupta, Heinrich Jiang, Erez Louidor, James Muller, Tamann Narayan, Serena Wang, Tao Zhu
Set functions predict a label from a permutation-invariant variable-size collection of feature vectors. We propose making set functions more understandable and regularized by capturing domain knowledge through shape constraints. We show how prior work in monotonic constraints can be adapted to set functions, and then propose two new shape constraints designed to generalize the conditioning role of weights in a weighted mean. We show how one can train standard functions and set functions that satisfy these shape constraints with a deep lattice network. We propose a nonlinear estimation strategy we call the semantic feature engine that uses set functions with the proposed shape constraints to estimate labels for compound sparse categorical features. Experiments on real-world data show the achieved accuracy is similar to deep sets or deep neural networks, but provides guarantees on the model behavior, which makes it easier to explain and debug.
****
Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints
Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, Seungil You

Classifiers can be trained with data-dependent constraints to satisfy fairness g
oals, reduce churn, achieve a targeted false positive rate, or other policy goal
s. We study the generalization performance for such constrained optimization pro
blems, in terms of how well the constraints are satisfied at evaluation time, gi
ven that they are satisfied at training time. To improve generalization, we fram
e the problem as a two-player game where one player optimizes the model paramete
rs on a training dataset, and the other player enforces the constraints on an in
dependent validation dataset. We build on recent work in two-player constrained
optimization to show that if one uses this two-dataset approach, then constraint
 generalization can be significantly improved. As we illustrate experimentally,
this approach works not only in theory, but also in practice.
****
Monge blunts Bayes: Hardness Results for Adversarial Training
Zac Cranko, Aditya Menon, Richard Nock, Cheng Soon Ong, Zhan Shi, Christian Wald
er
The last few years have seen a staggering number of empirical studies of the rob
ustness of neural networks in a model of adversarial perturbations of their inpu
ts. Most rely on an adversary which carries out local modifications within presc
ribed balls. None however has so far questioned the broader picture: how to fram
e a resource-bounded adversary so that it can be severely detrimental to learnin
g, a non-trivial problem which entails at a minimum the choice of loss and class
ifiers. We suggest a formal answer for losses that satisfy the minimal statistic
al requirement of being proper. We pin down a simple sufficient property for any
 given class of adversaries to be detrimental to learning, involving a central m
easure of “harmfulness” which generalizes the well-known class of integral proba
bility metrics. A key feature of our result is that it holds for all proper loss
es, and for a popular subset of these, the optimisation of this central measure
appears to be independent of the loss. When classifiers are Lipschitz – a now po
pular approach in adversarial training –, this optimisation resorts to optimal t
ransport to make a low-budget compression of class marginals. Toy experiments re
veal a finding recently separately observed: training against a sufficiently bud
geted adversary of this kind improves generalization.
****
Boosted Density Estimation Remastered
Zac Cranko, Richard Nock
There has recently been a steady increase in the number iterative approaches to
density estimation. However, an accompanying burst of formal convergence guarant
ees has not followed; all results pay the price of heavy assumptions which are o
ften unrealistic or hard to check. The Generative Adversarial Network (GAN) lite
rature — seemingly orthogonal to the aforementioned pursuit — has had the side e
ffect of a renewed interest in variational divergence minimisation (notably f-
GAN). We show how to combine this latter approach and the classical boosting the
ory in supervised learning to get the first density estimation algorithm that pr
ovably achieves geometric convergence under very weak assumptions. We do so by a
 trick allowing to combine classifiers as the sufficient statistics of an expone
ntial family. Our analysis includes an improved variational characterisation of
f-GAN.
****
Submodular Cost Submodular Cover with an Approximate Oracle
Victoria Crawford, Alan Kuhnle, My Thai
In this work, we study the Submodular Cost Submodular Cover problem, which is to
 minimize the submodular cost required to ensure that the submodular benefit fun
ction exceeds a given threshold. Existing approximation ratios for the greedy al
gorithm assume a value oracle to the benefit function. However, access to a valu
e oracle is not a realistic assumption for many applications of this problem, wh
ere the benefit function is difficult to compute. We present two incomparable ap
proximation ratios for this problem with an approximate value oracle and demonst
rate that the ratios take on empirically relevant values through a case study wi
th the Influence Threshold problem in online social networks.
****

Flexibly Fair Representation Learning by Disentanglement
Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, Richard Zemel
We consider the problem of learning representations that achieve group and subgroup fairness with respect to multiple sensitive attributes. Taking inspiration from the disentangled representation learning literature, we propose an algorithm for learning compact representations of datasets that are useful for reconstruction and prediction, but are also flexibly fair, meaning they can be easily modified at test time to achieve subgroup demographic parity with respect to multiple sensitive attributes and their conjunctions. We show empirically that the resulting encoder—which does not require the sensitive attributes for inference—allows for the adaptation of a single representation to a variety of fair classification tasks with new target labels and subgroup definitions.
****
Anytime Online-to-Batch, Optimism and Acceleration
Ashok Cutkosky
A standard way to obtain convergence guarantees in stochastic convex optimization is to run an online learning algorithm and then output the average of its iterates: the actual iterates of the online learning algorithm do not come with individual guarantees. We close this gap by introducing a black-box modification to any online learning algorithm whose iterates converge to the optimum in stochastic scenarios. We then consider the case of smooth losses, and show that combining our approach with optimistic online learning algorithms immediately yields a fast convergence rate of $O(L/T^{3/2}+\sigma/\sqrt{T})$ on L-smooth problems with σ^2 variance in the gradients. Finally, we provide a reduction that converts any adaptive online algorithm into one that obtains the optimal accelerated rate of $\tilde O(L/T^2 + \sigma/\sqrt{T})$, while still maintaining $\tilde O(1/\sqrt{T})$ convergence in the non-smooth setting. Importantly, our algorithms adapt to L and σ automatically: they do not need to know either to obtain these rates.
****
Matrix-Free Preconditioning in Online Learning
Ashok Cutkosky, Tamas Sarlos
We provide an online convex optimization algorithm with regret that interpolates between the regret of an algorithm using an optimal preconditioning matrix and one using a diagonal preconditioning matrix. Our regret bound is never worse than that obtained by diagonal preconditioning, and in certain setting even surpasses that of algorithms with full-matrix preconditioning. Importantly, our algorithm runs in the same time and space complexity as online gradient descent. Along the way we incorporate new techniques that mildly streamline and improve logarithmic factors in prior regret analyses. We conclude by benchmarking our algorithm on synthetic data and deep learning tasks.
****
Minimal Achievable Sufficient Statistic Learning
Milan Cvitkovic, Günther Koliander
We introduce Minimal Achievable Sufficient Statistic (MASS) Learning, a machine learning training objective for which the minima are minimal sufficient statistics with respect to a class of functions being optimized over (e.g., deep networks). In deriving MASS Learning, we also introduce Conserved Differential Information (CDI), an information-theoretic quantity that {—} unlike standard mutual information {—} can be usefully applied to deterministically-dependent continuous random variables like the input and output of a deep network. In a series of experiments, we show that deep networks trained with MASS Learning achieve competitive performance on supervised learning, regularization, and uncertainty quantification benchmarks.
****
Open Vocabulary Learning on Source Code with a Graph-Structured Cache
Milan Cvitkovic, Badal Singh, Animashree Anandkumar
Machine learning models that take computer program source code as input typically use Natural Language Processing (NLP) techniques. However, a major challenge i

s that code is written using an open, rapidly changing vocabulary due to, e.g., the coinage of new variable and method names. Reasoning over such a vocabulary is not something for which most NLP methods are designed. We introduce a Graph-Structured Cache to address this problem; this cache contains a node for each new word the model encounters with edges connecting each word to its occurrences in the code. We find that combining this graph-structured cache strategy with recent Graph-Neural-Network-based models for supervised learning on code improves the models' performance on a code completion task and a variable naming task — with over 100% relative improvement on the latter — at the cost of a moderate increase in computation time.
****
The Value Function Polytope in Reinforcement Learning
Robert Dadashi, Adrien Ali Taiga, Nicolas Le Roux, Dale Schuurmans, Marc G. Bellemare
We establish geometric and topological properties of the space of value functions in finite state-action Markov decision processes. Our main contribution is the characterization of the nature of its shape: a general polytope (Aigner et al., 2010). To demonstrate this result, we exhibit several properties of the structural relationship between policies and value functions including the line theorem, which shows that the value functions of policies constrained on all but one state describe a line segment. Finally, we use this novel perspective and introduce visualizations to enhance the understanding of the dynamics of reinforcement learning algorithms.
****
Bayesian Optimization Meets Bayesian Optimal Stopping
Zhongxiang Dai, Haibin Yu, Bryan Kian Hsiang Low, Patrick Jaillet
Bayesian optimization (BO) is a popular paradigm for optimizing the hyperparameters of machine learning (ML) models due to its sample efficiency. Many ML models require running an iterative training procedure (e.g., stochastic gradient descent). This motivates the question whether information available during the training process (e.g., validation accuracy after each epoch) can be exploited for improving the epoch efficiency of BO algorithms by early-stopping model training under hyperparameter settings that will end up under-performing and hence eliminating unnecessary training epochs. This paper proposes to unify BO (specifically, Gaussian process-upper confidence bound (GP-UCB)) with Bayesian optimal stopping (BO-BOS) to boost the epoch efficiency of BO. To achieve this, while GP-UCB is sample-efficient in the number of function evaluations, BOS complements it with epoch efficiency for each function evaluation by providing a principled optimal stopping mechanism for early stopping. BO-BOS preserves the (asymptotic) no-regret performance of GP-UCB using our specified choice of BOS parameters that is amenable to an elegant interpretation in terms of the exploration-exploitation trade-off. We empirically evaluate the performance of BO-BOS and demonstrate its generality in hyperparameter optimization of ML models and two other interesting applications.
****
Policy Certificates: Towards Accountable Reinforcement Learning
Christoph Dann, Lihong Li, Wei Wei, Emma Brunskill
The performance of a reinforcement learning algorithm can vary drastically during learning because of exploration. Existing algorithms provide little information about the quality of their current policy before executing it, and thus have limited use in high-stakes applications like healthcare. We address this lack of accountability by proposing that algorithms output policy certificates. These certificates bound the sub-optimality and return of the policy in the next episode, allowing humans to intervene when the certified quality is not satisfactory. We further introduce two new algorithms with certificates and present a new framework for theoretical analysis that guarantees the quality of their policies and certificates. For tabular MDPs, we show that computing certificates can even improve the sample-efficiency of optimism-based exploration. As a result, one of our algorithms is the first to achieve minimax-optimal PAC bounds up to lower-order terms, and this algorithm also matches (and in some settings slightly improves

upon) existing minimax regret bounds.
****
Learning Fast Algorithms for Linear Transforms Using Butterfly Factorizations
Tri Dao, Albert Gu, Matthew Eichhorn, Atri Rudra, Christopher Re
Fast linear transforms are ubiquitous in machine learning, including the discrete Fourier transform, discrete cosine transform, and other structured transformations such as convolutions. All of these transforms can be represented by dense matrix-vector multiplication, yet each has a specialized and highly efficient (subquadratic) algorithm. We ask to what extent hand-crafting these algorithms and implementations is necessary, what structural prior they encode, and how much knowledge is required to automatically learn a fast algorithm for a provided structured transform. Motivated by a characterization of fast matrix-vector multiplication as products of sparse matrices, we introduce a parameterization of divide-and-conquer methods that is capable of representing a large class of transforms. This generic formulation can automatically learn an efficient algorithm for many important transforms; for example, it recovers the $O(N \log N)$ Cooley-Tukey FFT algorithm to machine precision, for dimensions N up to 1024. Furthermore, our method can be incorporated as a lightweight replacement of generic matrices in machine learning pipelines to learn efficient and compressible transformations. On a standard task of compressing a single hidden-layer network, our method exceeds the classification accuracy of unconstrained matrices on CIFAR-10 by 3.9 points—the first time a structured approach has done so—with 4X faster inference speed and 40X fewer parameters.
****
A Kernel Theory of Modern Data Augmentation
Tri Dao, Albert Gu, Alexander Ratner, Virginia Smith, Chris De Sa, Christopher Re
Data augmentation, a technique in which a training set is expanded with class-preserving transformations, is ubiquitous in modern machine learning pipelines. In this paper, we seek to establish a theoretical framework for understanding data augmentation. We approach this from two directions: First, we provide a general model of augmentation as a Markov process, and show that kernels appear naturally with respect to this model, even when we do not employ kernel classification. Next, we analyze more directly the effect of augmentation on kernel classifiers, showing that data augmentation can be approximated by first-order feature averaging and second-order variance regularization components. These frameworks both serve to illustrate the ways in which data augmentation affects the downstream learning model, and the resulting analyses provide novel connections between prior work in invariant kernels, tangent propagation, and robust optimization. Finally, we provide several proof-of-concept applications showing that our theory can be useful for accelerating machine learning workflows, such as reducing the amount of computation needed to train using augmented data, and predicting the utility of a transformation prior to training.
****
TarMAC: Targeted Multi-Agent Communication
Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, Joelle Pineau
We propose a targeted communication architecture for multi-agent reinforcement learning, where agents learn both what messages to send and whom to address them to while performing cooperative tasks in partially-observable environments. This targeting behavior is learnt solely from downstream task-specific reward without any communication supervision. We additionally augment this with a multi-round communication approach where agents coordinate via multiple rounds of communication before taking actions in the environment. We evaluate our approach on a diverse set of cooperative multi-agent tasks, of varying difficulties, with varying number of agents, in a variety of environments ranging from 2D grid layouts of shapes and simulated traffic junctions to 3D indoor environments, and demonstrate the benefits of targeted and multi-round communication. Moreover, we show that the targeted communication strategies learned by agents are interpretable and intuitive. Finally, we show that our architecture can be easily extended to mixed

and competitive environments, leading to improved performance and sample comple
xity over recent state-of-the-art approaches.
****
Teaching a black-box learner
Sanjoy Dasgupta, Daniel Hsu, Stefanos Poulis, Xiaojin Zhu
One widely-studied model of teaching calls for a teacher to provide the minimal
set of labeled examples that uniquely specifies a target concept. The assumption
is that the teacher knows the learner’s hypothesis class, which is often not tr
ue of real-life teaching scenarios. We consider the problem of teaching a learne
r whose representation and hypothesis class are unknown—that is, the learner is
a black box. We show that a teacher who does not interact with the learner can d
o no better than providing random examples. We then prove, however, that with in
teraction, a teacher can efficiently find a set of teaching examples that is a p
rovably good approximation to the optimal set. As an illustration, we show how t
his scheme can be used to shrink training sets for any family of classifiers: th
at is, to find an approximately-minimal subset of training instances that yields
the same classifier as the entire set.
****
Stochastic Deep Networks
Gwendoline De Bie, Gabriel Peyré, Marco Cuturi
Machine learning is increasingly targeting areas where input data cannot be accu
rately described by a single vector, but can be modeled instead using the more f
lexible concept of random vectors, namely probability measures or more simply po
int clouds of varying cardinality. Using deep architectures on measures poses, h
owever, many challenging issues. Indeed, deep architectures are originally desig
ned to handle fixed-length vectors, or, using recursive mechanisms, ordered sequ
ences thereof. In sharp contrast, measures describe a varying number of weighted
observations with no particular order. We propose in this work a deep framework
designed to handle crucial aspects of measures, namely permutation invariances,
variations in weights and cardinality. Architectures derived from this pipeline
can (i) map measures to measures - using the concept of push-forward operators;
(ii) bridge the gap between measures and Euclidean spaces - through integration
steps. This allows to design discriminative networks (to classify or reduce the
dimensionality of input measures), generative architectures (to synthesize meas
ures) and recurrent pipelines (to predict measure dynamics). We provide a theore
tical analysis of these building blocks, review our architectures’ approximation
abilities and robustness w.r.t. perturbation, and try them on various discrimin
ative and generative tasks.
****
Learning-to-Learn Stochastic Gradient Descent with Biased Regularization
Giulia Denevi, Carlo Ciliberto, Riccardo Grazzi, Massimiliano Pontil
We study the problem of learning-to-learn: infer- ring a learning algorithm that
works well on a family of tasks sampled from an unknown distribution. As class
of algorithms we consider Stochastic Gradient Descent (SGD) on the true risk reg
ularized by the square euclidean distance from a bias vector. We present an aver
age excess risk bound for such a learning algorithm that quantifies the potentia
l benefit of using a bias vector with respect to the unbiased case. We then prop
ose a novel meta-algorithm to estimate the bias term online from a sequence of o
bserved tasks. The small memory footprint and low time complexity of our approac
h makes it appealing in practice while our theoretical analysis provides guarant
ees on the generalization properties of the meta-algorithm on new tasks. A key f
eature of our results is that, when the number of tasks grows and their vari- an
ce is relatively small, our learning-to-learn approach has a significant advanta
ge over learning each task in isolation by standard SGD without a bias term. Num
erical experiments demonstrate the effectiveness of our approach in practice.
****
A Multitask Multiple Kernel Learning Algorithm for Survival Analysis with Applic
ation to Cancer Biology
Onur Dereli, Ceyda O■uz, Mehmet Gönen
Predictive performance of machine learning algorithms on related problems can be

improved using multitask learning approaches. Rather than performing survival analysis on each data set to predict survival times of cancer patients, we developed a novel multitask approach based on multiple kernel learning (MKL). Our multitask MKL algorithm both works on multiple cancer data sets and integrates cancer-related pathways/gene sets into survival analysis. We tested our algorithm, which is named as Path2MSurv, on the Cancer Genome Atlas data sets analyzing gene expression profiles of 7,655 patients from 20 cancer types together with cancer-specific pathway/gene set collections. Path2MSurv obtained better or comparable predictive performance when benchmarked against random survival forest, survival support vector machine, and single-task variant of our algorithm. Path2MSurv has the ability to identify key pathways/gene sets in predicting survival times of patients from different cancer types.
****
Learning to Convolve: A Generalized Weight-Tying Approach
Nichita Diaconu, Daniel Worrall
Recent work (Cohen & Welling, 2016) has shown that generalizations of convolutions, based on group theory, provide powerful inductive biases for learning. In these generalizations, filters are not only translated but can also be rotated, flipped, etc. However, coming up with exact models of how to rotate a 3x3 filter on a square pixel-grid is difficult. In this paper, we learn how to transform filters for use in the group convolution, focussing on roto-translation. For this, we learn a filter basis and all rotated versions of that filter basis. Filters are then encoded by a set of rotation invariant coefficients. To rotate a filter, we switch the basis. We demonstrate we can produce feature maps with low sensitivity to input rotations, while achieving high performance on MNIST and CIFAR-10.
****
Sever: A Robust Meta-Algorithm for Stochastic Optimization
Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, Alistair Stewart
In high dimensions, most machine learning methods are brittle to even a small fraction of structured outliers. To address this, we introduce a new meta-algorithm that can take in a base learner such as least squares or stochastic gradient descent, and harden the learner to be resistant to outliers. Our method, Sever, possesses strong theoretical guarantees yet is also highly scalable – beyond running the base learner itself, it only requires computing the top singular vector of a certain n{\texttimes}d matrix. We apply Sever on a drug design dataset and a spam classification dataset, and find that in both cases it has substantially greater robustness than several baselines. On the spam dataset, with 1% corruptions, we achieved 7.4% test error, compared to 13.4%-20.5% for the baselines, and 3% error on the uncorrupted dataset. Similarly, on the drug design dataset, with 10% corruptions, we achieved 1.42 mean-squared error test error, compared to 1.51-2.33 for the baselines, and 1.23 error on the uncorrupted dataset.
****
Approximated Oracle Filter Pruning for Destructive CNN Width Optimization
Xiaohan Ding, Guiguang Ding, Yuchen Guo, Jungong Han, Chenggang Yan
It is not easy to design and run Convolutional Neural Networks (CNNs) due to: 1) finding the optimal number of filters (i.e., the width) at each layer is tricky, given an architecture; and 2) the computational intensity of CNNs impedes the deployment on computationally limited devices. Oracle Pruning is designed to remove the unimportant filters from a well-trained CNN, which estimates the filters’ importance by ablating them in turn and evaluating the model, thus delivers high accuracy but suffers from intolerable time complexity, and requires a given resulting width but cannot automatically find it. To address these problems, we propose Approximated Oracle Filter Pruning (AOFP), which keeps searching for the least important filters in a binary search manner, makes pruning attempts by masking out filters randomly, accumulates the resulting errors, and finetunes the model via a multi-path framework. As AOFP enables simultaneous pruning on multiple layers, we can prune an existing very deep CNN with acceptable time cost, negligible accuracy drop, and no heuristic knowledge, or re-design a model which exe

rts higher accuracy and faster inference.
****
Noisy Dual Principal Component Pursuit
Tianyu Ding, Zhihui Zhu, Tianjiao Ding, Yunchen Yang, Rene Vidal, Manolis Tsakiris, Daniel Robinson
Dual Principal Component Pursuit (DPCP) is a recently proposed non-convex optimization based method for learning subspaces of high relative dimension from noiseless datasets contaminated by as many outliers as the square of the number of inliers. Experimentally, DPCP has proved to be robust to noise and outperform the popular RANSAC on 3D vision tasks such as road plane detection and relative poses estimation from three views. This paper extends the global optimality and convergence theory of DPCP to the case of data corrupted by noise, and further demonstrates its robustness using synthetic and real data.
****
Finite-Time Analysis of Distributed TD(0) with Linear Function Approximation on Multi-Agent Reinforcement Learning
Thinh Doan, Siva Maguluri, Justin Romberg
We study the policy evaluation problem in multi-agent reinforcement learning. In this problem, a group of agents works cooperatively to evaluate the value function for the global discounted accumulative reward problem, which is composed of local rewards observed by the agents. Over a series of time steps, the agents act, get rewarded, update their local estimate of the value function, then communicate with their neighbors. The local update at each agent can be interpreted as a distributed consensus-based variant of the popular temporal difference learning algorithm TD(0). While distributed reinforcement learning algorithms have been presented in the literature, almost nothing is known about their convergence rate. Our main contribution is providing a finite-time analysis for the convergence of the distributed TD(0) algorithm. We do this when the communication network between the agents is time-varying in general. We obtain an explicit upper bound on the rate of convergence of this algorithm as a function of the network topology and the discount factor. Our results mirror what we would expect from using distributed stochastic gradient descent for solving convex optimization problems.
****
Trajectory-Based Off-Policy Deep Reinforcement Learning
Andreas Doerr, Michael Volpp, Marc Toussaint, Trimpe Sebastian, Christian Daniel
Policy gradient methods are powerful reinforcement learning algorithms and have been demonstrated to solve many complex tasks. However, these methods are also data-inefficient, afflicted with high variance gradient estimates, and frequently get stuck in local optima. This work addresses these weaknesses by combining recent improvements in the reuse of off-policy data and exploration in parameter space with deterministic behavioral policies. The resulting objective is amenable to standard neural network optimization strategies like stochastic gradient descent or stochastic gradient Hamiltonian Monte Carlo. Incorporation of previous rollouts via importance sampling greatly improves data-efficiency, whilst stochastic optimization schemes facilitate the escape from local optima. We evaluate the proposed approach on a series of continuous control benchmark tasks. The results show that the proposed algorithm is able to successfully and reliably learn solutions using fewer system interactions than standard policy gradient methods.
****
Generalized No Free Lunch Theorem for Adversarial Robustness
Elvis Dohmatob
This manuscript presents some new impossibility results on adversarial robustness in machine learning, a very important yet largely open problem. We show that if conditioned on a class label the data distribution satisfies the W_2 Talagrand transportation-cost inequality (for example, this condition is satisfied if the conditional distribution has density which is log-concave; is the uniform measure on a compact Riemannian manifold with positive Ricci curvature, any classifier can be adversarially fooled with high probability once the perturbations are slightly greater than the natural noise level in the problem. We call this resu

lt The Strong "No Free Lunch" Theorem as some recent results (Tsipras et al. 2018, Fawzi et al. 2018, etc.) on the subject can be immediately recovered as very particular cases. Our theoretical bounds are demonstrated on both simulated and real data (MNIST). We conclude the manuscript with some speculation on possible future research directions.
****
Width Provably Matters in Optimization for Deep Linear Neural Networks
Simon Du, Wei Hu
We prove that for an L-layer fully-connected linear neural network, if the width of every hidden layer is $\widetilde{\Omega}\left(L \cdot r \cdot d_{out} \cdot \kappa^3 \right)$, where r and κ are the rank and the condition number of the input data, and d_{out} is the output dimension, then gradient descent with Gaussian random initialization converges to a global minimum at a linear rate. The number of iterations to find an ϵ-suboptimal solution is $O(\kappa \log(\frac{1}{\epsilon}))$. Our polynomial upper bound on the total running time for wide deep linear networks and the $\exp\left(\Omega\left(L\right)\right)$ lower bound for narrow deep linear neural networks [Shamir, 2018] together demonstrate that wide layers are necessary for optimizing deep models.
****
Provably efficient RL with Rich Observations via Latent State Decoding
Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, John Langford
We study the exploration problem in episodic MDPs with rich observations generated from a small number of latent states. Under certain identifiability assumptions, we demonstrate how to estimate a mapping from the observations to latent states inductively through a sequence of regression and clustering steps—where previously decoded latent states provide labels for later regression problems—and use it to construct good exploration policies. We provide finite-sample guarantees on the quality of the learned state decoding function and exploration policies, and complement our theory with an empirical evaluation on a class of hard exploration problems. Our method exponentially improves over Q-learning with naïve exploration, even when Q-learning has cheating access to latent states.
****
Gradient Descent Finds Global Minima of Deep Neural Networks
Simon Du, Jason Lee, Haochuan Li, Liwei Wang, Xiyu Zhai
Gradient descent finds a global minimum in training deep neural networks despite the objective function being non-convex. The current paper proves gradient descent achieves zero training loss in polynomial time for a deep over-parameterized neural network with residual connections (ResNet). Our analysis relies on the particular structure of the Gram matrix induced by the neural network architecture. This structure allows us to show the Gram matrix is stable throughout the training process and this stability implies the global optimality of the gradient descent algorithm. We further extend our analysis to deep residual convolutional neural networks and obtain a similar convergence result.
****
Incorporating Grouping Information into Bayesian Decision Tree Ensembles
Junliang Du, Antonio Linero
We consider the problem of nonparametric regression in the high-dimensional setting in which $P \gg N$. We study the use of overlapping group structures to improve prediction and variable selection. These structures arise commonly when analyzing DNA microarray data, where genes can naturally be grouped according to genetic pathways. We incorporate overlapping group structure into a Bayesian additive regression trees model using a prior constructed so that, if a variable from some group is used to construct a split, this increases the probability that subsequent splits will use predictors from the same group. We refer to our model as an overlapping group Bayesian additive regression trees (OG-BART) model, and our prior on the splits an overlapping group Dirichlet (OG-Dirichlet) prior. Like the sparse group lasso, our prior encourages sparsity both within and between groups. We study the correlation structure of the prior, illustrate the proposed methodology on simulated data, and apply the methodology to gene expression data

to learn which genetic pathways are predictive of breast cancer tumor metastasis
.
****
Task-Agnostic Dynamics Priors for Deep Reinforcement Learning
Yilun Du, Karthic Narasimhan
While model-based deep reinforcement learning (RL) holds great promise for sampl
e efficiency and generalization, learning an accurate dynamics model is often ch
allenging and requires substantial interaction with the environment. A wide vari
ety of domains have dynamics that share common foundations like the laws of clas
sical mechanics, which are rarely exploited by existing algorithms. In fact, hum
ans continuously acquire and use such dynamics priors to easily adapt to operati
ng in new environments. In this work, we propose an approach to learn task-agnos
tic dynamics priors from videos and incorporate them into an RL agent. Our metho
d involves pre-training a frame predictor on task-agnostic physics videos to ini
tialize dynamics models (and fine-tune them) for unseen target environments. Our
 frame prediction architecture, SpatialNet, is designed specifically to capture
localized physical phenomena and interactions. Our approach allows for both fast
er policy learning and convergence to better policies, outperforming competitive
 approaches on several different environments. We also demonstrate that incorpor
ating this prior allows for more effective transfer between environments.
****
Optimal Auctions through Deep Learning
Paul Duetting, Zhe Feng, Harikrishna Narasimhan, David Parkes, Sai Srivatsa Ravi
ndranath
Designing an incentive compatible auction that maximizes expected revenue is an
intricate task. The single-item case was resolved in a seminal piece of work by
Myerson in 1981. Even after 30-40 years of intense research the problem remains
unsolved for seemingly simple multi-bidder, multi-item settings. In this work, w
e initiate the exploration of the use of tools from deep learning for the automa
ted design of optimal auctions. We model an auction as a multi-layer neural netw
ork, frame optimal auction design as a constrained learning problem, and show ho
w it can be solved using standard pipelines. We prove generalization bounds and
present extensive experiments, recovering essentially all known analytical solut
ions for multi-item settings, and obtaining novel mechanisms for settings in whi
ch the optimal mechanism is unknown.
****
Wasserstein of Wasserstein Loss for Learning Generative Models
Yonatan Dukler, Wuchen Li, Alex Lin, Guido Montufar
The Wasserstein distance serves as a loss function for unsupervised learning whi
ch depends on the choice of a ground metric on sample space. We propose to use t
he Wasserstein distance itself as the ground metric on the sample space of image
s. This ground metric is known as an effective distance for image retrieval, tha
t correlates with human perception. We derive the Wasserstein ground metric on p
ixel space and define a Riemannian Wasserstein gradient penalty to be used in th
e Wasserstein Generative Adversarial Network (WGAN) framework. The new gradient
penalty is computed efficiently via convolutions on the L^2 gradients with neg
ligible additional computational cost. The new formulation is more robust to the
 natural variability of the data and provides for a more continuous discriminato
r in sample space.
****
Learning interpretable continuous-time models of latent stochastic dynamical sys
tems
Lea Duncker, Gergo Bohner, Julien Boussard, Maneesh Sahani
We develop an approach to learn an interpretable semi-parametric model of a late
nt continuous-time stochastic dynamical system, assuming noisy high-dimensional
outputs sampled at uneven times. The dynamics are described by a nonlinear stoch
astic differential equation (SDE) driven by a Wiener process, with a drift evolu
tion function drawn from a Gaussian process (GP) conditioned on a set of learnt
fixed points and corresponding local Jacobian matrices. This form yields a flexi
ble nonparametric model of the dynamics, with a representation corresponding dir

ectly to the interpretable portraits routinely employed in the study of nonlinea r dynamical systems. The learning algorithm combines inference of continuous lat ent paths underlying observed data with a sparse variational description of the dynamical process. We demonstrate our approach on simulated data from different nonlinear dynamical systems.
****
Autoregressive Energy Machines
Charlie Nash, Conor Durkan
Neural density estimators are flexible families of parametric models which have seen widespread use in unsupervised machine learning in recent years. Maximum-li kelihood training typically dictates that these models be constrained to specify an explicit density. However, this limitation can be overcome by instead using a neural network to specify an energy function, or unnormalized density, which c an subsequently be normalized to obtain a valid distribution. The challenge with this approach lies in accurately estimating the normalizing constant of the hig h-dimensional energy function. We propose the Autoregressive Energy Machine, an energy-based model which simultaneously learns an unnormalized density and compu tes an importance-sampling estimate of the normalizing constant for each conditi onal in an autoregressive decomposition. The Autoregressive Energy Machine achie ves state-of-the-art performance on a suite of density-estimation tasks.
****
Band-limited Training and Inference for Convolutional Neural Networks
Adam Dziedzic, John Paparrizos, Sanjay Krishnan, Aaron Elmore, Michael Franklin
The convolutional layers are core building blocks of neural network architecture s. In general, a convolutional filter applies to the entire frequency spectrum o f the input data. We explore artificially constraining the frequency spectra of these filters and data, called band-limiting, during training. The frequency dom ain constraints apply to both the feed-forward and back-propagation steps. Exper imentally, we observe that Convolutional Neural Networks (CNNs) are resilient to this compression scheme and results suggest that CNNs learn to leverage lower-f requency components. In particular, we found: (1) band-limited training can effe ctively control the resource usage (GPU and memory); (2) models trained with ban d-limited layers retain high prediction accuracy; and (3) requires no modificati on to existing training algorithms or neural network architectures to use unlike other compression schemes.
****
Imitating Latent Policies from Observation
Ashley Edwards, Himanshu Sahni, Yannick Schroecker, Charles Isbell
In this paper, we describe a novel approach to imitation learning that infers la tent policies directly from state observations. We introduce a method that chara cterizes the causal effects of latent actions on observations while simultaneous ly predicting their likelihood. We then outline an action alignment procedure th at leverages a small amount of environment interactions to determine a mapping b etween the latent and real-world actions. We show that this corrected labeling c an be used for imitating the observed behavior, even though no expert actions ar e given. We evaluate our approach within classic control environments and a plat form game and demonstrate that it performs better than standard approaches. Code for this work is available at https://github.com/ashedwards/ILPO.
****
Semi-Cyclic Stochastic Gradient Descent
Hubert Eichner, Tomer Koren, Brendan Mcmahan, Nathan Srebro, Kunal Talwar
We consider convex SGD updates with a block-cyclic structure, i.e., where each c ycle consists of a small number of blocks, each with many samples from a possibl y different, block-specific, distribution. This situation arises, e.g., in Feder ated Learning where the mobile devices available for updates at different times during the day have different characteristics. We show that such block-cyclic st ructure can significantly deteriorate the performance of SGD, but propose a simp le approach that allows prediction with the same guarantees as for i.i.d., non-c yclic, sampling.
****

GDPP: Learning Diverse Generations using Determinantal Point Processes
Mohamed Elfeki, Camille Couprie, Morgane Riviere, Mohamed Elhoseiny
Generative models have proven to be an outstanding tool for representing high-dimensional probability distributions and generating realistic looking images. An essential characteristic of generative models is their ability to produce multi-modal outputs. However, while training, they are often susceptible to mode collapse, that is models are limited in mapping input noise to only a few modes of the true data distribution. In this work, we draw inspiration from Determinantal Point Process (DPP) to propose an unsupervised penalty loss that alleviates mode collapse while producing higher quality samples. DPP is an elegant probabilistic measure used to model negative correlations within a subset and hence quantify its diversity. We use DPP kernel to model the diversity in real data as well as in synthetic data. Then, we devise an objective term that encourages generator to synthesize data with a similar diversity to real data. In contrast to previous state-of-the-art generative models that tend to use additional trainable parameters or complex training paradigms, our method does not change the original training scheme. Embedded in an adversarial training and variational autoencoder, our Generative DPP approach shows a consistent resistance to mode-collapse on a wide-variety of synthetic data and natural image datasets including MNIST, CIFAR10, and CelebA, while outperforming state-of-the-art methods for data-efficiency, generation quality, and convergence-time whereas being 5.8x faster than its closest competitor.
****
Sequential Facility Location: Approximate Submodularity and Greedy Algorithm
Ehsan Elhamifar
We develop and analyze a novel utility function and a fast optimization algorithm for subset selection in sequential data that incorporates the dynamic model of data. We propose a cardinality-constrained sequential facility location function that finds a fixed number of representatives, where the sequence of representatives is compatible with the dynamic model and well encodes the data. As maximizing this new objective function is NP-hard, we develop a fast greedy algorithm based on submodular maximization. Unlike the conventional facility location, the computation of the marginal gain in our case cannot be done by operations on each item independently. We exploit the sequential structure of the problem and develop an efficient dynamic programming-based algorithm that computes the marginal gain exactly. We investigate conditions on the dynamic model, under which our utility function is (ϵ-approximately) submodualr, hence, the greedy algorithm comes with performance guarantees. By experiments on synthetic data and the problem of procedure learning from instructional videos, we show that our framework significantly improves the computational time, achieves better objective function values and obtains more coherent summaries.
****
Improved Convergence for ℓ_1 and ℓ_∞ Regression via Iteratively Reweighted Least Squares
Alina Ene, Adrian Vladu
The iteratively reweighted least squares method (IRLS) is a popular technique used in practice for solving regression problems. Various versions of this method have been proposed, but their theoretical analyses failed to capture the good practical performance. In this paper we propose a simple and natural version of IRLS for solving ℓ_∞ and ℓ_1 regression, which provably converges to a $(1+\epsilon)$-approximate solution in $O(m^{1/3}\log(1/\epsilon)/\epsilon^{2/3} + \log m/\epsilon^2)$ iterations, where m is the number of rows of the input matrix. Interestingly, this running time is independent of the conditioning of the input, and the dominant term of the running time depends sublinearly in ϵ^{-1}, which is atypical for the optimization of non-smooth functions. This improves upon the more complex algorithms of Chin et al. (ITCS ’12), and Christiano et al. (STOC ’11) by a factor of at least $1/\epsilon^2$, and yields a truly efficient natural algorithm for the slime mold dynamics (Straszak-Vishnoi, SODA ’16, ITCS ’16, ITCS ’17).
****

Exploring the Landscape of Spatial Robustness
Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, Aleksander Madry
The study of adversarial robustness has so far largely focused on perturbations bound in ℓ_p-norms. However, state-of-the-art models turn out to be also vulnerable to other, more natural classes of perturbations such as translations and rotations. In this work, we thoroughly investigate the vulnerability of neural network–based classifiers to rotations and translations. While data augmentation offers relatively small robustness, we use ideas from robust optimization and test-time input aggregation to significantly improve robustness. Finally we find that, in contrast to the ℓ_p-norm case, first-order methods cannot reliably find worst-case perturbations. This highlights spatial robustness as a fundamentally different setting requiring additional study.
****
Cross-Domain 3D Equivariant Image Embeddings
Carlos Esteves, Avneesh Sud, Zhengyi Luo, Kostas Daniilidis, Ameesh Makadia
Spherical convolutional networks have been introduced recently as tools to learn powerful feature representations of 3D shapes. Spherical CNNs are equivariant to 3D rotations making them ideally suited to applications where 3D data may be observed in arbitrary orientations. In this paper we learn 2D image embeddings with a similar equivariant structure: embedding the image of a 3D object should commute with rotations of the object. We introduce a cross-domain embedding from 2D images into a spherical CNN latent space. This embedding encodes images with 3D shape properties and is equivariant to 3D rotations of the observed object. The model is supervised only by target embeddings obtained from a spherical CNN pretrained for 3D shape classification. We show that learning a rich embedding for images with appropriate geometric structure is sufficient for tackling varied applications, such as relative pose estimation and novel view synthesis, without requiring additional task-specific supervision.
****
On the Connection Between Adversarial Robustness and Saliency Map Interpretability
Christian Etmann, Sebastian Lunz, Peter Maass, Carola Schoenlieb
Recent studies on the adversarial vulnerability of neural networks have shown that models trained to be more robust to adversarial attacks exhibit more interpretable saliency maps than their non-robust counterparts. We aim to quantify this behaviour by considering the alignment between input image and saliency map. We hypothesize that as the distance to the decision boundary grows, so does the alignment. This connection is strictly true in the case of linear models. We confirm these theoretical findings with experiments based on models trained with a local Lipschitz regularization and identify where the nonlinear nature of neural networks weakens the relation.
****
Non-monotone Submodular Maximization with Nearly Optimal Adaptivity and Query Complexity
Matthew Fahrbach, Vahab Mirrokni, Morteza Zadimoghaddam
Submodular maximization is a general optimization problem with a wide range of applications in machine learning (e.g., active learning, clustering, and feature selection). In large-scale optimization, the parallel running time of an algorithm is governed by its adaptivity, which measures the number of sequential rounds needed if the algorithm can execute polynomially-many independent oracle queries in parallel. While low adaptivity is ideal, it is not sufficient for an algorithm to be efficient in practice—there are many applications of distributed submodular optimization where the number of function evaluations becomes prohibitively expensive. Motivated by these applications, we study the adaptivity and query complexity of submodular maximization. In this paper, we give the first constant-factor approximation algorithm for maximizing a non-monotone submodular function subject to a cardinality constraint k that runs in $O(\log(n))$ adaptive rounds and makes $O(n \log(k))$ oracle queries in expectation. In our empirical study, we use three real-world applications to compare our algorithm with several benchmarks for non-monotone submodular maximization. The results demonstrate that

 our algorithm finds competitive solutions using significantly fewer rounds and
queries.
****
Multi-Frequency Vector Diffusion Maps
Yifeng Fan, Zhizhen Zhao
We introduce multi-frequency vector diffusion maps (MFVDM), a new framework for
organizing and analyzing high dimensional data sets. The new method is a mathema
tical and algorithmic generalization of vector diffusion maps (VDM) and other no
n-linear dimensionality reduction methods. The idea of MFVDM is to incorporates
multiple unitary irreducible representations of the alignment group which introd
uces robustness to noise. We illustrate the efficacy of MFVDM on synthetic and c
ryo-EM image datasets, achieving better nearest neighbors search and alignment e
stimation than other baselines as VDM and diffusion maps (DM), especially on ext
remely noisy data.
****
Stable-Predictive Optimistic Counterfactual Regret Minimization
Gabriele Farina, Christian Kroer, Noam Brown, Tuomas Sandholm
The CFR framework has been a powerful tool for solving large-scale extensive-for
m games in practice. However, the theoretical rate at which past CFR-based algor
ithms converge to the Nash equilibrium is on the order of $O(T^{-1/2})$, where $
T$ is the number of iterations. In contrast, first-order methods can be used to
achieve a $O(T^{-1})$ dependence on iterations, yet these methods have been less
 successful in practice. In this work we present the first CFR variant that brea
ks the square-root dependence on iterations. By combining and extending recent a
dvances on predictive and stable regret minimizers for the matrix-game setting w
e show that it is possible to leverage “optimistic” regret minimizers to achieve
 a $O(T^{-3/4})$ convergence rate within CFR. This is achieved by introducing a
new notion of stable-predictivity, and by setting the stability of each counterf
actual regret minimizer relative to its location in the decision tree. Experimen
ts show that this method is faster than the original CFR algorithm, although not
 as fast as newer variants, in spite of their worst-case $O(T^{-1/2})$ dependenc
e on iterations.
****
Regret Circuits: Composability of Regret Minimizers
Gabriele Farina, Christian Kroer, Tuomas Sandholm
Regret minimization is a powerful tool for solving large-scale problems; it was
recently used in breakthrough results for large-scale extensive-form game solvin
g. This was achieved by composing simplex regret minimizers into an overall regr
et-minimization framework for extensive-form game strategy spaces. In this paper
 we study the general composability of regret minimizers. We derive a calculus f
or constructing regret minimizers for composite convex sets that are obtained fr
om convexity-preserving operations on simpler convex sets. We show that local re
gret minimizers for the simpler sets can be combined with additional regret mini
mizers into an aggregate regret minimizer for the composite set. As one applicat
ion, we show that the CFR framework can be constructed easily from our framework
. We also show ways to include curtailing (constraining) operations into our fra
mework. For one, they enable the construction of CFR generalization for extensiv
e-form games with general convex strategy constraints that can cut across decisi
on points.
****
Dead-ends and Secure Exploration in Reinforcement Learning
Mehdi Fatemi, Shikhar Sharma, Harm Van Seijen, Samira Ebrahimi Kahou
Many interesting applications of reinforcement learning (RL) involve MDPs that i
nclude numerous “dead-end" states. Upon reaching a dead-end state, the agent con
tinues to interact with the environment in a dead-end trajectory before reaching
 an undesired terminal state, regardless of whatever actions are chosen. The sit
uation is even worse when existence of many dead-end states is coupled with dist
ant positive rewards from any initial state (we term this as Bridge Effect). Hen
ce, conventional exploration techniques often incur prohibitively many training
steps before convergence. To deal with the bridge effect, we propose a condition

for exploration, called security. We next establish formal results that translate the security condition into the learning problem of an auxiliary value function. This new value function is used to cap “any" given exploration policy and is guaranteed to make it secure. As a special case, we use this theory and introduce secure random-walk. We next extend our results to the deep RL settings by identifying and addressing two main challenges that arise. Finally, we empirically compare secure random-walk with standard benchmarks in two sets of experiments including the Atari game of Montezuma’s Revenge.
****
Invariant-Equivariant Representation Learning for Multi-Class Data
Ilya Feige
Representations learnt through deep neural networks tend to be highly informative, but opaque in terms of what information they learn to encode. We introduce an approach to probabilistic modelling that learns to represent data with two separate deep representations: an invariant representation that encodes the information of the class from which the data belongs, and an equivariant representation that encodes the symmetry transformation defining the particular data point within the class manifold (equivariant in the sense that the representation varies naturally with symmetry transformations). This approach is based primarily on the strategic routing of data through the two latent variables, and thus is conceptually transparent, easy to implement, and in-principle generally applicable to any data comprised of discrete classes of continuous distributions (e.g. objects in images, topics in language, individuals in behavioural data). We demonstrate qualitatively compelling representation learning and competitive quantitative performance, in both supervised and semi-supervised settings, versus comparable modelling approaches in the literature with little fine tuning.
****
The advantages of multiple classes for reducing overfitting from test set reuse
Vitaly Feldman, Roy Frostig, Moritz Hardt
Excessive reuse of holdout data can lead to overfitting. However, there is little concrete evidence of significant overfitting due to holdout reuse in popular multiclass benchmarks today. Known results show that, in the worst-case, revealing the accuracy of k adaptively chosen classifiers on a data set of size n allows to create a classifier with bias of $\Theta(\sqrt{k/n})$ for any binary prediction problem. We show a new upper bound of $\tilde O(\max\{\sqrt{k\log(n)/(mn)}, k/n\})$ on the worst-case bias that any attack can achieve in a prediction problem with m classes. Moreover, we present an efficient attack that achieve a bias of $\Omega(\sqrt{k/(m^2 n)})$ and improves on previous work for the binary setting ($m=2$). We also present an inefficient attack that achieves a bias of $\tilde\Omega(k/n)$. Complementing our theoretical work, we give new practical attacks to stress-test multiclass benchmarks by aiming to create as large a bias as possible with a given number of queries. Our experiments show that the additional uncertainty of prediction with a large number of classes indeed mitigates the effect of our best attacks.
****
Decentralized Exploration in Multi-Armed Bandits
Raphael Feraud, Reda Alami, Romain Laroche
We consider the decentralized exploration problem: a set of players collaborate to identify the best arm by asynchronously interacting with the same stochastic environment. The objective is to insure privacy in the best arm identification problem between asynchronous, collaborative, and thrifty players. In the context of a digital service, we advocate that this decentralized approach allows a good balance between conflicting interests: the providers optimize their services, while protecting privacy of users and saving resources. We define the privacy level as the amount of information an adversary could infer by intercepting all the messages concerning a single user. We provide a generic algorithm DECENTRALIZED ELIMINATION, which uses any best arm identification algorithm as a subroutine. We prove that this algorithm insures privacy, with a low communication cost, and that in comparison to the lower bound of the best arm identification problem, its sample complexity suffers from a penalty depending on the inverse of the prob

ability of the most frequent players. Then, thanks to the genericity of the approach, we extend the proposed algorithm to the non-stationary bandits. Finally, experiments illustrate and complete the analysis.
****
Almost surely constrained convex optimization
Olivier Fercoq, Ahmet Alacaoglu, Ion Necoara, Volkan Cevher
We propose a stochastic gradient framework for solving stochastic composite convex optimization problems with (possibly) infinite number of linear inclusion constraints that need to be satisfied almost surely. We use smoothing and homotopy techniques to handle constraints without the need for matrix-valued projections. We show for our stochastic gradient algorithm $\mathcal{O}(\log(k)/\sqrt{k})$ convergence rate for general convex objectives and $\mathcal{O}(\log(k)/k)$ convergence rate for restricted strongly convex objectives. These rates are known to be optimal up to logarithmic factor, even without constraints. We conduct numerical experiments on basis pursuit, hard margin support vector machines and portfolio optimization problems and show that our algorithm achieves state-of-the-art practical performance.
****
Online Meta-Learning
Chelsea Finn, Aravind Rajeswaran, Sham Kakade, Sergey Levine
A central capability of intelligent systems is the ability to continuously build upon previous experiences to speed up and enhance learning of new tasks. Two distinct research paradigms have studied this question. Meta-learning views this problem as learning a prior over model parameters that is amenable for fast adaptation on a new task, but typically assumes the tasks are available together as a batch. In contrast, online (regret based) learning considers a setting where tasks are revealed one after the other, but conventionally trains a single model without task-specific adaptation. This work introduces an online meta-learning setting, which merges ideas from both paradigms to better capture the spirit and practice of continual lifelong learning. We propose the follow the meta leader (FTML) algorithm which extends the MAML algorithm to this setting. Theoretically, this work provides an O(log T) regret guarantee with one additional higher order smoothness assumption (in comparison to the standard online setting). Our experimental evaluation on three different large-scale problems suggest that the proposed algorithm significantly outperforms alternatives based on traditional online learning approaches.
****
DL2: Training and Querying Neural Networks with Logic
Marc Fischer, Mislav Balunovic, Dana Drachsler-Cohen, Timon Gehr, Ce Zhang, Martin Vechev
We present DL2, a system for training and querying neural networks with logical constraints. Using DL2, one can declaratively specify domain knowledge constraints to be enforced during training, as well as pose queries on the model to find inputs that satisfy a set of constraints. DL2 works by translating logical constraints into a loss function with desirable mathematical properties. The loss is then minimized with standard gradient-based methods. We evaluate DL2 by training networks with interesting constraints in unsupervised, semi-supervised and supervised settings. Our experimental evaluation demonstrates that DL2 is more expressive than prior approaches combining logic and neural networks, and its loss functions are better suited for optimization. Further, we show that for a number of queries, DL2 can find the desired inputs in seconds (even for large models such as ResNet-50 on ImageNet).
****
Bayesian Action Decoder for Deep Multi-Agent Reinforcement Learning
Jakob Foerster, Francis Song, Edward Hughes, Neil Burch, Iain Dunning, Shimon Whiteson, Matthew Botvinick, Michael Bowling
When observing the actions of others, humans make inferences about why they acted as they did, and what this implies about the world; humans also use the fact that their actions will be interpreted in this manner, allowing them to act informatively and thereby communicate efficiently with others. Although learning algo

rithms have recently achieved superhuman performance in a number of two-player, zero-sum games, scalable multi-agent reinforcement learning algorithms that can discover effective strategies and conventions in complex, partially observable settings have proven elusive. We present the Bayesian action decoder (BAD), a new multi-agent learning method that uses an approximate Bayesian update to obtain a public belief that conditions on the actions taken by all agents in the environment. BAD introduces a new Markov decision process, the public belief MDP, in which the action space consists of all deterministic partial policies, and exploits the fact that an agent acting only on this public belief state can still learn to use its private information if the action space is augmented to be over all partial policies mapping private information into environment actions. The Bayesian update is closely related to the theory of mind reasoning that humans carry out when observing others' actions. We first validate BAD on a proof-of-principle two-step matrix game, where it outperforms policy gradient methods; we then evaluate BAD on the challenging, cooperative partial-information card game Hanabi, where, in the two-player setting, it surpasses all previously published learning and hand-coded approaches, establishing a new state of the art.

****

Scalable Nonparametric Sampling from Multimodal Posteriors with the Posterior Bootstrap

Edwin Fong, Simon Lyddon, Chris Holmes

Increasingly complex datasets pose a number of challenges for Bayesian inference. Conventional posterior sampling based on Markov chain Monte Carlo can be too computationally intensive, is serial in nature and mixes poorly between posterior modes. Furthermore, all models are misspecified, which brings into question the validity of the conventional Bayesian update. We present a scalable Bayesian nonparametric learning routine that enables posterior sampling through the optimization of suitably randomized objective functions. A Dirichlet process prior on the unknown data distribution accounts for model misspecification, and admits an embarrassingly parallel posterior bootstrap algorithm that generates independent and exact samples from the nonparametric posterior distribution. Our method is particularly adept at sampling from multimodal posterior distributions via a random restart mechanism, and we demonstrate this on Gaussian mixture model and sparse logistic regression examples.

****

On discriminative learning of prediction uncertainty

Vojtech Franc, Daniel Prusa

In classification with a reject option, the classifier is allowed in uncertain cases to abstain from prediction. The classical cost based model of an optimal classifier with a reject option requires the cost of rejection to be defined explicitly. An alternative bounded-improvement model, avoiding the notion of the reject cost, seeks for a classifier with a guaranteed selective risk and maximal cover. We prove that both models share the same class of optimal strategies, and we provide an explicit relation between the reject cost and the target risk being the parameters of the two models. An optimal rejection strategy for both models is based on thresholding the conditional risk defined by posterior probabilities which are usually unavailable. We propose a discriminative algorithm learning an uncertainty function which preserves ordering of the input space induced by the conditional risk, and hence can be used to construct optimal rejection strategies.

****

Learning Discrete Structures for Graph Neural Networks

Luca Franceschi, Mathias Niepert, Massimiliano Pontil, Xiao He

Graph neural networks (GNNs) are a popular class of machine learning models that have been successfully applied to a range of problems. Their major advantage lies in their ability to explicitly incorporate a sparse and discrete dependency structure between data points. Unfortunately, GNNs can only be used when such a graph-structure is available. In practice, however, real-world graphs are often noisy and incomplete or might not be available at all. With this work, we propose to jointly learn the graph structure and the parameters of graph convolutional

networks (GCNs) by approximately solving a bilevel program that learns a discret e probability distribution on the edges of the graph. This allows one to apply G CNs not only in scenarios where the given graph is incomplete or corrupted but a lso in those where a graph is not available. We conduct a series of experiments that analyze the behavior of the proposed method and demonstrate that it outperf orms related methods by a significant margin.
****
Distributional Multivariate Policy Evaluation and Exploration with the Bellman G AN
Dror Freirich, Tzahi Shimkin, Ron Meir, Aviv Tamar
The recently proposed distributional approach to reinforcement learning (DiRL) i s centered on learning the distribution of the reward-to-go, often referred to a s the value distribution. In this work, we show that the distributional Bellman equation, which drives DiRL methods, is equivalent to a generative adversarial n etwork (GAN) model. In this formulation, DiRL can be seen as learning a deep gen erative model of the value distribution, driven by the discrepancy between the d istribution of the current value, and the distribution of the sum of current rew ard and next value. We use this insight to propose a GAN-based approach to DiRL, which leverages the strengths of GANs in learning distributions of high dimensi onal data. In particular, we show that our GAN approach can be used for DiRL wit h multivariate rewards, an important setting which cannot be tackled with prior methods. The multivariate setting also allows us to unify learning the distribut ion of values and state transitions, and we exploit this idea to devise a novel exploration method that is driven by the discrepancy in estimating both values a nd states.
****
Approximating Orthogonal Matrices with Effective Givens Factorization
Thomas Frerix, Joan Bruna
We analyze effective approximation of unitary matrices. In our formulation, a un itary matrix is represented as a product of rotations in two-dimensional subspac es, so-called Givens rotations. Instead of the quadratic dimension dependence wh en applying a dense matrix, applying such an approximation scales with the numbe r factors, each of which can be implemented efficiently. Consequently, in settin gs where an approximation is once computed and then applied many times, such a r epresentation becomes advantageous. Although effective Givens factorization is n ot possible for generic unitary operators, we show that minimizing a sparsity-in ducing objective with a coordinate descent algorithm on the unitary group yields good factorizations for structured matrices. Canonical applications of such a s etup are orthogonal basis transforms. We demonstrate numerical results of approx imating the graph Fourier transform, which is the matrix obtained when diagonali zing a graph Laplacian.
****
Fast and Flexible Inference of Joint Distributions from their Marginals
Charlie Frogner, Tomaso Poggio
Across the social sciences and elsewhere, practitioners frequently have to reaso n about relationships between random variables, despite lacking joint observatio ns of the variables. This is sometimes called an "ecological" inference; given s amples from the marginal distributions of the variables, one attempts to infer t heir joint distribution. The problem is inherently ill-posed, yet only a few mod els have been proposed for bringing prior information into the problem, often re lying on restrictive or unrealistic assumptions and lacking a unified approach. In this paper, we treat the inference problem generally and propose a unified cl ass of models that encompasses some of those previously proposed while including many new ones. Previous work has relied on either relaxation or approximate inf erence via MCMC, with the latter known to mix prohibitively slowly for this type of problem. Here we instead give a single exact inference algorithm that works for the entire model class via an efficient fixed point iteration called Dykstra ’s method. We investigate empirically both the computational cost of our algorit hm and the accuracy of the new models on real datasets, showing favorable perfor mance in both cases and illustrating the impact of increased flexibility in mode

ling enabled by this work.
****
Analyzing and Improving Representations with the Soft Nearest Neighbor Loss
Nicholas Frosst, Nicolas Papernot, Geoffrey Hinton
We explore and expand the Soft Nearest Neighbor Loss to measure the entanglement of class manifolds in representation space: i.e., how close pairs of points from the same class are relative to pairs of points from different classes. We demonstrate several use cases of the loss. As an analytical tool, it provides insights into the evolution of class similarity structures during learning. Surprisingly, we find that maximizing the entanglement of representations of different classes in the hidden layers is beneficial for discrimination in the final layer, possibly because it encourages representations to identify class-independent similarity structures. Maximizing the soft nearest neighbor loss in the hidden layers leads not only to better-calibrated estimates of uncertainty on outlier data but also marginally improved generalization. Data that is not from the training distribution can be recognized by observing that in the hidden layers, it has fewer than the normal number of neighbors from the predicted class.
****
Diagnosing Bottlenecks in Deep Q-learning Algorithms
Justin Fu, Aviral Kumar, Matthew Soh, Sergey Levine
Q-learning methods are a common class of algorithms used in reinforcement learning (RL). However, their behavior with function approximation, especially with neural networks, is poorly understood theoretically and empirically. In this work, we aim to experimentally investigate potential issues in Q-learning, by means of a "unit testing" framework where we can utilize oracles to disentangle sources of error. Specifically, we investigate questions related to function approximation, sampling error and nonstationarity, and where available, verify if trends found in oracle settings hold true with deep RL methods. We find that large neural network architectures have many benefits with regards to learning stability; offer several practical compensations for overfitting; and develop a novel sampling method based on explicitly compensating for function approximation error that yields fair improvement on high-dimensional continuous control domains.
****
MetricGAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement
Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, Shou-De Lin
Adversarial loss in a conditional generative adversarial network (GAN) is not designed to directly optimize evaluation metrics of a target task, and thus, may not always guide the generator in a GAN to generate data with improved metric scores. To overcome this issue, we propose a novel MetricGAN approach with an aim to optimize the generator with respect to one or multiple evaluation metrics. Moreover, based on MetricGAN, the metric scores of the generated data can also be arbitrarily specified by users. We tested the proposed MetricGAN on a speech enhancement task, which is particularly suitable to verify the proposed approach because there are multiple metrics measuring different aspects of speech signals. Moreover, these metrics are generally complex and could not be fully optimized by Lp or conventional adversarial losses.
****
Beyond Adaptive Submodularity: Approximation Guarantees of Greedy Policy with Adaptive Submodularity Ratio
Kaito Fujii, Shinsaku Sakaue
We propose a new concept named adaptive submodularity ratio to study the greedy policy for sequential decision making. While the greedy policy is known to perform well for a wide variety of adaptive stochastic optimization problems in practice, its theoretical properties have been analyzed only for a limited class of problems. We narrow the gap between theory and practice by using adaptive submodularity ratio, which enables us to prove approximation guarantees of the greedy policy for a substantially wider class of problems. Examples of newly analyzed problems include important applications such as adaptive influence maximization and adaptive feature selection. Our adaptive submodularity ratio also provides bou

nds of adaptivity gaps. Experiments confirm that the greedy policy performs well with the applications being considered compared to standard heuristics.
****
Off-Policy Deep Reinforcement Learning without Exploration
Scott Fujimoto, David Meger, Doina Precup
Many practical applications of reinforcement learning constrain agents to learn from a fixed batch of data which has already been gathered, without offering further possibility for data collection. In this paper, we demonstrate that due to errors introduced by extrapolation, standard off-policy deep reinforcement learning algorithms, such as DQN and DDPG, are incapable of learning with data uncorrelated to the distribution under the current policy, making them ineffective for this fixed batch setting. We introduce a novel class of off-policy algorithms, batch-constrained reinforcement learning, which restricts the action space in order to force the agent towards behaving close to on-policy with respect to a subset of the given data. We present the first continuous control deep reinforcement learning algorithm which can learn effectively from arbitrary, fixed batch data, and empirically demonstrate the quality of its behavior in several tasks.
****
Transfer Learning for Related Reinforcement Learning Tasks via Image-to-Image Translation
Shani Gamrian, Yoav Goldberg
Despite the remarkable success of Deep RL in learning control policies from raw pixels, the resulting models do not generalize. We demonstrate that a trained agent fails completely when facing small visual changes, and that fine-tuning—the common transfer learning paradigm—fails to adapt to these changes, to the extent that it is faster to re-train the model from scratch. We show that by separating the visual transfer task from the control policy we achieve substantially better sample efficiency and transfer behavior, allowing an agent trained on the source task to transfer well to the target tasks. The visual mapping from the target to the source domain is performed using unaligned GANs, resulting in a control policy that can be further improved using imitation learning from imperfect demonstrations. We demonstrate the approach on synthetic visual variants of the Breakout game, as well as on transfer between subsequent levels of Road Fighter, a Nintendo car-driving game. A visualization of our approach can be seen in \url{https://youtu.be/4mnkzYyXMn4} and \url{https://youtu.be/KCGTrQi6Ogo}.
****
Breaking the Softmax Bottleneck via Learnable Monotonic Pointwise Non-linearities
Octavian Ganea, Sylvain Gelly, Gary Becigneul, Aliaksei Severyn
The Softmax function on top of a final linear layer is the de facto method to output probability distributions in neural networks. In many applications such as language models or text generation, this model has to produce distributions over large output vocabularies. Recently, this has been shown to have limited representational capacity due to its connection with the rank bottleneck in matrix factorization. However, little is known about the limitations of Linear-Softmax for quantities of practical interest such as cross entropy or mode estimation, a direction that we explore here. As an efficient and effective solution to alleviate this issue, we propose to learn parametric monotonic functions on top of the logits. We theoretically investigate the rank increasing capabilities of such monotonic functions. Empirically, our method improves in two different quality metrics over the traditional Linear-Softmax layer in synthetic and real language model experiments, adding little time or memory overhead, while being comparable to the more computationally expensive mixture of Softmaxes.
****
Graph U-Nets
Hongyang Gao, Shuiwang Ji
We consider the problem of representation learning for graph data. Convolutional neural networks can naturally operate on images, but have significant challenges in dealing with graph data. Given images are special cases of graphs with nodes lie on 2D lattices, graph embedding tasks have a natural correspondence with i

mage pixel-wise prediction tasks such as segmentation. While encoder-decoder arc
hitectures like U-Nets have been successfully applied on many image pixel-wise p
rediction tasks, similar methods are lacking for graph data. This is due to the
fact that pooling and up-sampling operations are not natural on graph data. To a
ddress these challenges, we propose novel graph pooling (gPool) and unpooling (g
Unpool) operations in this work. The gPool layer adaptively selects some nodes t
o form a smaller graph based on their scalar projection values on a trainable pr
ojection vector. We further propose the gUnpool layer as the inverse operation o
f the gPool layer. The gUnpool layer restores the graph into its original struct
ure using the position information of nodes selected in the corresponding gPool
layer. Based on our proposed gPool and gUnpool layers, we develop an encoder-dec
oder model on graph, known as the graph U-Nets. Our experimental results on node
 classification and graph classification tasks demonstrate that our methods achi
eve consistently better performance than previous models.
****
Deep Generative Learning via Variational Gradient Flow
Yuan Gao, Yuling Jiao, Yang Wang, Yao Wang, Can Yang, Shunkang Zhang
We propose a framework to learn deep generative models via \textbf{V}ariational
\textbf{Gr}adient Fl\textbf{ow} (VGrow) on probability spaces. The evolving dist
ribution that asymptotically converges to the target distribution is governed by
 a vector field, which is the negative gradient of the first variation of the $f
$-divergence between them. We prove that the evolving distribution coincides wit
h the pushforward distribution through the infinitesimal time composition of res
idual maps that are perturbations of the identity map along the vector field. Th
e vector field depends on the density ratio of the pushforward distribution and
the target distribution, which can be consistently learned from a binary classif
ication problem. Connections of our proposed VGrow method with other popular met
hods, such as VAE, GAN and flow-based methods, have been established in this fra
mework, gaining new insights of deep generative learning. We also evaluated seve
ral commonly used divergences, including Kullback-Leibler, Jensen-Shannon, Jeffr
eys divergences as well as our newly discovered “logD” divergence which serves a
s the objective function of the logD-trick GAN. Experimental results on benchmar
k datasets demonstrate that VGrow can generate high-fidelity images in a stable
and efficient manner, achieving competitive performance with state-of-the-art GA
Ns.
****
Rate Distortion For Model Compression:From Theory To Practice
Weihao Gao, Yu-Han Liu, Chong Wang, Sewoong Oh
The enormous size of modern deep neural net-works makes it challenging to deploy
 those models in memory and communication limited scenarios. Thus, compressing a
 trained model without a significant loss in performance has become an increasin
gly important task. Tremendous advances has been made recently, where the main t
echnical building blocks are pruning, quantization, and low-rank factorization.
In this paper, we propose principled approaches to improve upon the common heuri
stics used in those building blocks, by studying the fundamental limit for model
 compression via the rate distortion theory. We prove a lower bound for the rate
 distortion function for model compression and prove its achievability for linea
r models. Although this achievable compression scheme is intractable in practice
, this analysis motivates a novel objective function for model compression, whic
h can be used to improve classes of model compressor such as pruning or quantiza
tion. Theoretically, we prove that the proposed scheme is optimal for compressin
g one-hidden-layer ReLU neural networks. Empirically,we show that the proposed s
cheme improves upon the baseline in the compression-accuracy tradeoff.
****
Demystifying Dropout
Hongchang Gao, Jian Pei, Heng Huang
Dropout is a popular technique to train large-scale deep neural networks to alle
viate the overfitting problem. To disclose the underlying reasons for its gain,
numerous works have tried to explain it from different perspectives. In this pap
er, unlike existing works, we explore it from a new perspective to provide new i

nsight into this line of research. In detail, we disentangle the forward and bac
kward pass of dropout. Then, we find that these two passes need different levels
 of noise to improve the generalization performance of deep neural networks. Bas
ed on this observation, we propose the augmented dropout which employs different
 dropping strategies in the forward and backward pass. Experimental results have
 verified the effectiveness of our proposed method.
****
Geometric Scattering for Graph Data Analysis
Feng Gao, Guy Wolf, Matthew Hirn
We explore the generalization of scattering transforms from traditional (e.g., i
mage or audio) signals to graph data, analogous to the generalization of ConvNet
s in geometric deep learning, and the utility of extracted graph features in gra
ph data analysis. In particular, we focus on the capacity of these features to r
etain informative variability and relations in the data (e.g., between individua
l graphs, or in aggregate), while relating our construction to previous theoreti
cal results that establish the stability of similar transforms to families of gr
aph deformations. We demonstrate the application of our geometric scattering fea
tures in graph classification of social network data, and in data exploration of
 biochemistry data.
****
Multi-Frequency Phase Synchronization
Tingran Gao, Zhizhen Zhao
We propose a novel formulation for phase synchronization—the statistical problem
 of jointly estimating alignment angles from noisy pairwise comparisons—as a non
convex optimization problem that enforces consistency among the pairwise compari
sons in multiple frequency channels. Inspired by harmonic retrieval in signal pr
ocessing, we develop a simple yet efficient two-stage algorithm that leverages t
he multi-frequency information. We demonstrate in theory and practice that the p
roposed algorithm significantly outperforms state-of-the-art phase synchronizati
on algorithms, at a mild computational costs incurred by using the extra frequen
cy channels. We also extend our algorithmic framework to general synchronization
 problems over compact Lie groups.
****
Optimal Mini-Batch and Step Sizes for SAGA
Nidham Gazagnadou, Robert Gower, Joseph Salmon
Recently it has been shown that the step sizes of a family of variance reduced g
radient methods called the JacSketch methods depend on the expected smoothness c
onstant. In particular, if this expected smoothness constant could be calculated
 a priori, then one could safely set much larger step sizes which would result i
n a much faster convergence rate. We fill in this gap, and provide simple closed
 form expressions for the expected smoothness constant and careful numerical exp
eriments verifying these bounds. Using these bounds, and since the SAGA algorith
m is part of this JacSketch family, we suggest a new standard practice for setti
ng the step and mini-batch sizes for SAGA that are competitive with a numerical
grid search. Furthermore, we can now show that the total complexity of the SAGA
algorithm decreases linearly in the mini-batch size up to a pre-defined value: t
he optimal mini-batch size. This is a rare result in the stochastic variance red
uced literature, only previously shown for the Katyusha algorithm. Finally we co
njecture that this is the case for many other stochastic variance reduced method
s and that our bounds and analysis of the expected smoothness constant is key to
 extending these results.
****
SelectiveNet: A Deep Neural Network with an Integrated Reject Option
Yonatan Geifman, Ran El-Yaniv
We consider the problem of selective prediction (also known as reject option) in
 deep neural networks, and introduce SelectiveNet, a deep neural architecture wi
th an integrated reject option. Existing rejection mechanisms are based mostly o
n a threshold over the prediction confidence of a pre-trained network. In contra
st, SelectiveNet is trained to optimize both classification (or regression) and
rejection simultaneously, end-to-end. The result is a deep neural network that i

s optimized over the covered domain. In our experiments, we show a consistently improved risk-coverage trade-off over several well-known classification and regression datasets, thus reaching new state-of-the-art results for deep selective classification.
****

A Theory of Regularized Markov Decision Processes
Matthieu Geist, Bruno Scherrer, Olivier Pietquin
Many recent successful (deep) reinforcement learning algorithms make use of regularization, generally based on entropy or Kullback-Leibler divergence. We propose a general theory of regularized Markov Decision Processes that generalizes these approaches in two directions: we consider a larger class of regularizers, and we consider the general modified policy iteration approach, encompassing both policy iteration and value iteration. The core building blocks of this theory are a notion of regularized Bellman operator and the Legendre-Fenchel transform, a classical tool of convex optimization. This approach allows for error propagation analyses of general algorithmic schemes of which (possibly variants of) classical algorithms such as Trust Region Policy Optimization, Soft Q-learning, Stochastic Actor Critic or Dynamic Policy Programming are special cases. This also draws connections to proximal convex optimization, especially to Mirror Descent.
****

DeepMDP: Learning Continuous Latent Space Models for Representation Learning
Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, Marc G. Bellemare
Many reinforcement learning (RL) tasks provide the agent with high-dimensional observations that can be simplified into low-dimensional continuous states. To formalize this process, we introduce the concept of a \texit{DeepMDP}, a parameterized latent space model that is trained via the minimization of two tractable latent space losses: prediction of rewards and prediction of the distribution over next latent states. We show that the optimization of these objectives guarantees (1) the quality of the embedding function as a representation of the state space and (2) the quality of the DeepMDP as a model of the environment. Our theoretical findings are substantiated by the experimental result that a trained DeepMDP recovers the latent structure underlying high-dimensional observations on a synthetic environment. Finally, we show that learning a DeepMDP as an auxiliary task in the Atari 2600 domain leads to large performance improvements over model-free RL.
****

Partially Linear Additive Gaussian Graphical Models
Sinong Geng, Minhao Yan, Mladen Kolar, Sanmi Koyejo
We propose a partially linear additive Gaussian graphical model (PLA-GGM) for the estimation of associations between random variables distorted by observed confounders. Model parameters are estimated using an L_1-regularized maximal pseudo-profile likelihood estimator (MaPPLE) for which we prove a $\sqrt{n}$-sparsistency. Importantly, our approach avoids parametric constraints on the effects of confounders on the estimated graphical model structure. Empirically, the PLA-GGM is applied to both synthetic and real-world datasets, demonstrating superior performance compared to competing methods.
****

Learning and Data Selection in Big Datasets
Hossein Shokri Ghadikolaei, Hadi Ghauch, Carlo Fischione, Mikael Skoglund
Finding a dataset of minimal cardinality to characterize the optimal parameters of a model is of paramount importance in machine learning and distributed optimization over a network. This paper investigates the compressibility of large datasets. More specifically, we propose a framework that jointly learns the input-output mapping as well as the most representative samples of the dataset (sufficient dataset). Our analytical results show that the cardinality of the sufficient dataset increases sub-linearly with respect to the original dataset size. Numerical evaluations of real datasets reveal a large compressibility, up to 95%, without a noticeable drop in the learnability performance, measured by the generalization error.
****

Improved Parallel Algorithms for Density-Based Network Clustering
Mohsen Ghaffari, Silvio Lattanzi, Slobodan Mitrovi■
Clustering large-scale networks is a central topic in unsupervised learning with
 many applications in machine learning and data mining. A classic approach to cl
uster a network is to identify regions of high edge density, which in the litera
ture is captured by two fundamental problems: the densest subgraph and the k-c
ore decomposition problems. We design massively parallel computation (MPC) algor
ithms for these problems that are considerably faster than prior work. In the ca
se of k-core decomposition, our work improves exponentially on the algorithm p
rovided by Esfandiari et al. (ICML’18). Compared to the prior work on densest su
bgraph presented by Bahmani et al. (VLDB’12, ’14), our result requires quadratic
ally fewer MPC rounds. We complement our analysis with an experimental scalabili
ty analysis of our techniques.
****
Recursive Sketches for Modular Deep Learning
Badih Ghazi, Rina Panigrahy, Joshua Wang
We present a mechanism to compute a sketch (succinct summary) of how a complex m
odular deep network processes its inputs. The sketch summarizes essential inform
ation about the inputs and outputs of the network and can be used to quickly ide
ntify key components and summary statistics of the inputs. Furthermore, the sket
ch is recursive and can be unrolled to identify sub-components of these componen
ts and so forth, capturing a potentially complicated DAG structure. These sketch
es erase gracefully; even if we erase a fraction of the sketch at random, the re
mainder still retains the “high-weight” information present in the original sket
ch. The sketches can also be organized in a repository to implicitly form a “kno
wledge graph”; it is possible to quickly retrieve sketches in the repository tha
t are related to a sketch of interest; arranged in this fashion, the sketches ca
n also be used to learn emerging concepts by looking for new clusters in sketch
space. Finally, in the scenario where we want to learn a ground truth deep netwo
rk, we show that augmenting input/output pairs with these sketches can theoretic
ally make it easier to do so.
****
An Instability in Variational Inference for Topic Models
Behrooz Ghorbani, Hamid Javadi, Andrea Montanari
Naive mean field variational methods are the state of-the-art approach to infere
nce in topic modeling. We show that these methods suffer from an instability tha
t can produce misleading conclusions. Namely, for certain regimes of the model p
arameters, variational inference outputs a non-trivial decomposition into topics
. However -for the same parameter values- the data contain no actual information
 about the true topic decomposition, and the output of the algorithm is uncorrel
ated with it. In particular, the estimated posterior mean is wrong, and estimate
d credible regions do not achieve the nominal coverage. We discuss how this inst
ability is remedied by more accurate mean field approximations.
****
An Investigation into Neural Net Optimization via Hessian Eigenvalue Density
Behrooz Ghorbani, Shankar Krishnan, Ying Xiao
To understand the dynamics of training in deep neural networks, we study the evo
lution of the Hessian eigenvalue density throughout the optimization process. In
 non-batch normalized networks, we observe the rapid appearance of large isolate
d eigenvalues in the spectrum, along with a surprising concentration of the grad
ient in the corresponding eigenspaces. In a batch normalized network, these two
effects are almost absent. We give a theoretical rationale to partially explain
these phenomena. As part of this work, we adapt advanced tools from numerical li
near algebra that allow scalable and accurate estimation of the entire Hessian s
pectrum of ImageNet-scale neural networks; this technique may be of independent
interest in other applications.
****
Data Shapley: Equitable Valuation of Data for Machine Learning
Amirata Ghorbani, James Zou
As data becomes the fuel driving technological and economic growth, a fundamenta

l challenge is how to quantify the value of data in algorithmic predictions and decisions. For example, in healthcare and consumer markets, it has been suggeste d that individuals should be compensated for the data that they generate, but it is not clear what is an equitable valuation for individual data. In this work, we develop a principled framework to address data valuation in the context of su pervised machine learning. Given a learning algorithm trained on n data points to produce a predictor, we propose data Shapley as a metric to quantify the val ue of each training datum to the predictor performance. Data Shapley uniquely sa tisfies several natural properties of equitable data valuation. We develop Monte Carlo and gradient-based methods to efficiently estimate data Shapley values in practical settings where complex learning algorithms, including neural networks , are trained on large datasets. In addition to being equitable, extensive exper iments across biomedical, image and synthetic data demonstrate that data Shapley has several other benefits: 1) it is more powerful than the popular leave-one-o ut or leverage score in providing insight on what data is more valuable for a gi ven learning task; 2) low Shapley value data effectively capture outliers and co rruptions; 3) high Shapley value data inform what type of new data to acquire to improve the predictor.
****
Efficient Dictionary Learning with Gradient Descent
Dar Gilboa, Sam Buchanan, John Wright
Randomly initialized first-order optimization algorithms are the method of choic e for solving many high-dimensional nonconvex problems in machine learning, yet general theoretical guarantees cannot rule out convergence to critical points of poor objective value. For some highly structured nonconvex problems however, th e success of gradient descent can be understood by studying the geometry of the objective. We study one such problem – complete orthogonal dictionary learning, and provide converge guarantees for randomly initialized gradient descent to the neighborhood of a global optimum. The resulting rates scale as low order polyno mials in the dimension even though the objective possesses an exponential number of saddle points. This efficient convergence can be viewed as a consequence of negative curvature normal to the stable manifolds associated with saddle points, and we provide evidence that this feature is shared by other nonconvex problems of importance as well.
****
A Tree-Based Method for Fast Repeated Sampling of Determinantal Point Processes
Jennifer Gillenwater, Alex Kulesza, Zelda Mariet, Sergei Vassilvtiskii
It is often desirable in recommender systems and other information retrieval app lications to provide diverse results, and determinantal point processes (DPPs) h ave become a popular way to capture the trade-off between the quality of individ ual results and the diversity of the overall set. However, sampling from a DPP i s inherently expensive: if the underlying collection contains N items, then gene rating each DPP sample requires time linear in N following a one-time preprocess ing phase. Additionally, results often need to be personalized to a user, but st andard approaches to personalization invalidate the preprocessing, making person alized samples especially expensive. In this work we address both of these short comings. First, we propose a new algorithm for generating DPP samples in time lo garithmic in N, following a slightly more expensive preprocessing phase. We then extend the algorithm to support arbitrary query-time feature weights, allowing us to generate samples customized to individual users while still retaining loga rithmic runtime; experiments show our approach runs over 300 times faster than t raditional DPP sampling on collections of 100,000 items for samples of size 10.
****
Learning to Groove with Inverse Sequence Transformations
Jon Gillick, Adam Roberts, Jesse Engel, Douglas Eck, David Bamman
We explore models for translating abstract musical ideas (scores, rhythms) into expressive performances using seq2seq and recurrent variational information bott leneck (VIB) models. Though seq2seq models usually require painstakingly aligned corpora, we show that it is possible to adapt an approach from the Generative A dversarial Network (GAN) literature (e.g. Pix2Pix, Vid2Vid) to sequences, creati

ng large volumes of paired data by performing simple transformations and training generative models to plausibly invert these transformations. Music, and drumming in particular, provides a strong test case for this approach because many common transformations (quantization, removing voices) have clear semantics, and learning to invert them has real-world applications. Focusing on the case of drum set players, we create and release a new dataset for this purpose, containing over 13 hours of recordings by professional drummers aligned with fine-grained timing and dynamics information. We also explore some of the creative potential of these models, demonstrating improvements on state-of-the-art methods for Humanization (instantiating a performance from a musical score).

****

Adversarial Examples Are a Natural Consequence of Test Error in Noise

Justin Gilmer, Nicolas Ford, Nicholas Carlini, Ekin Cubuk

Over the last few years, the phenomenon of adversarial examples — maliciously constructed inputs that fool trained machine learning models — has captured the attention of the research community, especially when restricted to small modifications of a correctly handled input. Less surprisingly, image classifiers also lack human-level performance on randomly corrupted images, such as images with additive Gaussian noise. In this paper we provide both empirical and theoretical evidence that these are two manifestations of the same underlying phenomenon, and therefore the adversarial robustness and corruption robustness research programs are closely related. This suggests that improving adversarial robustness should go hand in hand with improving performance in the presence of more general and realistic image corruptions. This yields a computationally tractable evaluation metric for defenses to consider: test error in noisy image distributions.

****

Discovering Conditionally Salient Features with Statistical Guarantees

Jaime Roquero Gimenez, James Zou

The goal of feature selection is to identify important features that are relevant to explain a outcome variable. Most of the work in this domain has focused on identifying globally relevant features, which are features that are related to the outcome using evidence across the entire dataset. We study a more fine-grained statistical problem: conditional feature selection, where a feature may be relevant depending on the values of the other features. For example in genetic association studies, variant A could be associated with the phenotype in the entire dataset, but conditioned on variant B being present it might be independent of the phenotype. In this sense, variant A is globally relevant, but conditioned on B it is no longer locally relevant in that region of the feature space. We present a generalization of the knockoff procedure that performs conditional feature selection while controlling a generalization of the false discovery rate (FDR) to the conditional setting. By exploiting the feature/response model-free framework of the knockoffs, the quality of the statistical FDR guarantee is not degraded even when we perform conditional feature selections. We implement this method and present an algorithm that automatically partitions the feature space such that it enhances the differences between selected sets in different regions, and validate the statistical theoretical results with experiments.

****

Estimating Information Flow in Deep Neural Networks

Ziv Goldfeld, Ewout Van Den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, Yury Polyanskiy

We study the estimation of the mutual information I(X;T_ℓ) between the input X to a deep neural network (DNN) and the output vector T_ℓ of its ℓ-th hidden layer (an “internal representation”). Focusing on feedforward networks with fixed weights and noisy internal representations, we develop a rigorous framework for accurate estimation of I(X;T_ℓ). By relating I(X;T_ℓ) to information transmission over additive white Gaussian noise channels, we reveal that compression, i.e. reduction in I(X;T_ℓ) over the course of training, is driven by progressive geometric clustering of the representations of samples from the same class. Experimental results verify this connection. Finally, we shift focus to purely deterministic DNNs, where I(X;T_ℓ) is provably vacuous, a

nd show that nevertheless, these models also cluster inputs belonging to the same class. The binning-based approximation of I(X;T_ℓ) employed in past works to measure compression is identified as a measure of clustering, thus clarifying that these experiments were in fact tracking the same clustering phenomenon. Leveraging the clustering perspective, we provide new evidence that compression and generalization may not be causally related and discuss potential future research ideas.
****
Amortized Monte Carlo Integration
Adam Golinski, Frank Wood, Tom Rainforth
Current approaches to amortizing Bayesian inference focus solely on approximating the posterior distribution. Typically, this approximation is, in turn, used to calculate expectations for one or more target functions{—}a computational pipeline which is inefficient when the target function(s) are known upfront. In this paper, we address this inefficiency by introducing AMCI, a method for amortizing Monte Carlo integration directly. AMCI operates similarly to amortized inference but produces three distinct amortized proposals, each tailored to a different component of the overall expectation calculation. At runtime, samples are produced separately from each amortized proposal, before being combined to an overall estimate of the expectation. We show that while existing approaches are fundamentally limited in the level of accuracy they can achieve, AMCI can theoretically produce arbitrarily small errors for any integrable target function using only a single sample from each proposal at runtime. We further show that it is able to empirically outperform the theoretically optimal selfnormalized importance sampler on a number of example problems. Furthermore, AMCI allows not only for amortizing over datasets but also amortizing over target functions.
****
Online Algorithms for Rent-Or-Buy with Expert Advice
Sreenivas Gollapudi, Debmalya Panigrahi
We study the use of predictions by multiple experts (such as machine learning algorithms) to improve the performance of online algorithms. In particular, we consider the classical rent-or-buy problem (also called ski rental), and obtain algorithms that provably improve their performance over the adversarial scenario by using these predictions. We also prove matching lower bounds to show that our algorithms are the best possible, and perform experiments to empirically validate their performance in practice
****
The information-theoretic value of unlabeled data in semi-supervised learning
Alexander Golovnev, David Pal, Balazs Szorenyi
We quantify the separation between the numbers of labeled examples required to learn in two settings: Settings with and without the knowledge of the distribution of the unlabeled data. More specifically, we prove a separation by $\Theta(\log n)$ multiplicative factor for the class of projections over the Boolean hypercube of dimension n. We prove that there is no separation for the class of all functions on domain of any size. Learning with the knowledge of the distribution (a.k.a. fixed-distribution learning) can be viewed as an idealized scenario of semi-supervised learning where the number of unlabeled data points is so great that the unlabeled distribution is known exactly. For this reason, we call the separation the value of unlabeled data.
****
Efficient Training of BERT by Progressively Stacking
Linyuan Gong, Di He, Zhuohan Li, Tao Qin, Liwei Wang, Tieyan Liu
Unsupervised pre-training is popularly used in natural language processing. By designing proper unsupervised prediction tasks, a deep neural network can be trained and shown to be effective in many downstream tasks. As the data is usually adequate, the model for pre-training is generally huge and contains millions of parameters. Therefore, the training efficiency becomes a critical issue even when using high-performance hardware. In this paper, we explore an efficient training method for the state-of-the-art bidirectional Transformer (BERT) model. By visualizing the self-attention distribution of different layers at different positi

ons in a well-trained BERT model, we find that in most layers, the self-attention distribution will concentrate locally around its position and the start-of-sentence token. Motivating from this, we propose the stacking algorithm to transfer knowledge from a shallow model to a deep model; then we apply stacking progressively to accelerate BERT training. The experimental results showed that the models trained by our training strategy achieve similar performance to models trained from scratch, but our algorithm is much faster.
****
Quantile Stein Variational Gradient Descent for Batch Bayesian Optimization
Chengyue Gong, Jian Peng, Qiang Liu
Batch Bayesian optimization has been shown to be an efficient and successful approach for black-box function optimization, especially when the evaluation of cost function is highly expensive but can be efficiently parallelized. In this paper, we introduce a novel variational framework for batch query optimization, based on the argument that the query batch should be selected to have both high diversity and good worst case performance. This motivates us to introduce a variational objective that combines a quantile-based risk measure (for worst case performance) and entropy regularization (for enforcing diversity). We derive a gradient-based particle-based algorithm for solving our quantile-based variational objective, which generalizes Stein variational gradient descent (SVGD). We evaluate our method on a number of real-world applications and show that it consistently outperforms other recent state-of-the-art batch Bayesian optimization methods. Extensive experimental results indicate that our method achieves better or comparable performance, compared to the existing methods.
****
Obtaining Fairness using Optimal Transport Theory
Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, Jean-Michel Loubes
In the fair classification setup, we recast the links between fairness and predictability in terms of probability metrics. We analyze repair methods based on mapping conditional distributions to the Wasserstein barycenter. We propose a Random Repair which yields a tradeoff between minimal information loss and a certain amount of fairness.
****
Combining parametric and nonparametric models for off-policy evaluation
Omer Gottesman, Yao Liu, Scott Sussex, Emma Brunskill, Finale Doshi-Velez
We consider a model-based approach to perform batch off-policy evaluation in reinforcement learning. Our method takes a mixture-of-experts approach to combine parametric and non-parametric models of the environment such that the final value estimate has the least expected error. We do so by first estimating the local accuracy of each model and then using a planner to select which model to use at every time step as to minimize the return error estimate along entire trajectories. Across a variety of domains, our mixture-based approach outperforms the individual models alone as well as state-of-the-art importance sampling-based estimators.
****
Counterfactual Visual Explanations
Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, Stefan Lee
In this work, we develop a technique to produce counterfactual visual explanations. Given a ‘query’ image I for which a vision system predicts class c, a counterfactual visual explanation identifies how I could change such that the system would output a different specified class $c’$. To do this, we select a ‘distractor’ image $I’$ that the system predicts as class $c’$ and identify spatial regions in I and $I’$ such that replacing the identified region in I with the identified region in $I’$ would push the system towards classifying I as $c’$. We apply our approach to multiple image classification datasets generating qualitative results showcasing the interpretability and discriminativeness of our counterfactual explanations. To explore the effectiveness of our explanations in teaching humans, we present machine teaching experiments for the task of fine-grained bird classification. We find that users trained to distinguish bird species fare better when given access to counterfactual explanations in addition to t

raining examples.
****
Adaptive Sensor Placement for Continuous Spaces
James Grant, Alexis Boukouvalas, Ryan-Rhys Griffiths, David Leslie, Sattar Vakil i, Enrique Munoz De Cote
We consider the problem of adaptively placing sensors along an interval to detec t stochastically-generated events. We present a new formulation of the problem a s a continuum-armed bandit problem with feedback in the form of partial observat ions of realisations of an inhomogeneous Poisson process. We design a solution m ethod by combining Thompson sampling with nonparametric inference via increasing ly granular Bayesian histograms and derive an $\tilde{O}(T^{2/3})$ bound on the Bayesian regret in T rounds. This is coupled with the design of an efficent op timisation approach to select actions in polynomial time. In simulations we demo nstrate our approach to have substantially lower and less variable regret than c ompetitor algorithms.
****
A Statistical Investigation of Long Memory in Language and Music
Alexander Greaves-Tunnell, Zaid Harchaoui
Representation and learning of long-range dependencies is a central challenge co nfronted in modern applications of machine learning to sequence data. Yet despit e the prominence of this issue, the basic problem of measuring long-range depend ence, either in a given data source or as represented in a trained deep model, r emains largely limited to heuristic tools. We contribute a statistical framework for investigating long-range dependence in current applications of deep sequenc e modeling, drawing on the well-developed theory of long memory stochastic proce sses. This framework yields testable implications concerning the relationship be tween long memory in real-world data and its learned representation in a deep le arning architecture, which are explored through a semiparametric framework adapt ed to the high-dimensional setting.
****
Automatic Posterior Transformation for Likelihood-Free Inference
David Greenberg, Marcel Nonnenmacher, Jakob Macke
How can one perform Bayesian inference on stochastic simulators with intractable likelihoods? A recent approach is to learn the posterior from adaptively propos ed simulations using neural network-based conditional density estimators. Howeve r, existing methods are limited to a narrow range of proposal distributions or r equire importance weighting that can limit performance in practice. Here we pres ent automatic posterior transformation (APT), a new sequential neural posterior estimation method for simulation-based inference. APT can modify the posterior e stimate using arbitrary, dynamically updated proposals, and is compatible with p owerful flow-based density estimators. It is more flexible, scalable and efficie nt than previous simulation-based inference techniques. APT can operate directly on high-dimensional time series and image data, opening up new applications for likelihood-free inference.
****
Learning to Optimize Multigrid PDE Solvers
Daniel Greenfeld, Meirav Galun, Ronen Basri, Irad Yavneh, Ron Kimmel
Constructing fast numerical solvers for partial differential equations (PDEs) is crucial for many scientific disciplines. A leading technique for solving large-scale PDEs is using multigrid methods. At the core of a multigrid solver is the prolongation matrix, which relates between different scales of the problem. This matrix is strongly problem-dependent, and its optimal construction is critical to the efficiency of the solver. In practice, however, devising multigrid algori thms for new problems often poses formidable challenges. In this paper we propos e a framework for learning multigrid solvers. Our method learns a (single) mappi ng from discretized PDEs to prolongation operators for a broad class of 2D diffu sion problems. We train a neural network once for the entire class of PDEs, usin g an efficient and unsupervised loss function. Our tests demonstrate improved co nvergence rates compared to the widely used Black-Box multigrid scheme, suggesti ng that our method successfully learned rules for constructing prolongation matr

ices.
****
Multi-Object Representation Learning with Iterative Variational Inference
Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, Alexander Lerchner
Human perception is structured around objects which form the basis for our higher-level cognition and impressive systematic generalization abilities. Yet most work on representation learning focuses on feature learning without even considering multiple objects, or treats segmentation as an (often supervised) preprocessing step. Instead, we argue for the importance of learning to segment and represent objects jointly. We demonstrate that, starting from the simple assumption that a scene is composed of multiple entities, it is possible to learn to segment images into interpretable objects with disentangled representations. Our method learns – without supervision – to inpaint occluded parts, and extrapolates to scenes with more objects and to unseen objects with novel feature combinations. We also show that, due to the use of iterative variational inference, our system is able to learn multi-modal posteriors for ambiguous inputs and extends naturally to sequences.
****
Graphite: Iterative Generative Modeling of Graphs
Aditya Grover, Aaron Zweig, Stefano Ermon
Graphs are a fundamental abstraction for modeling relational data. However, graphs are discrete and combinatorial in nature, and learning representations suitable for machine learning tasks poses statistical and computational challenges. In this work, we propose Graphite, an algorithmic framework for unsupervised learning of representations over nodes in large graphs using deep latent variable generative models. Our model parameterizes variational autoencoders (VAE) with graph neural networks, and uses a novel iterative graph refinement strategy inspired by low-rank approximations for decoding. On a wide variety of synthetic and benchmark datasets, Graphite outperforms competing approaches for the tasks of density estimation, link prediction, and node classification. Finally, we derive a theoretical connection between message passing in graph neural networks and mean-field variational inference.
****
Fast Algorithm for Generalized Multinomial Models with Ranking Data
Jiaqi Gu, Guosheng Yin
We develop a framework of generalized multinomial models, which includes both the popular Plackett–Luce model and Bradley–Terry model as special cases. From a theoretical perspective, we prove that the maximum likelihood estimator (MLE) under generalized multinomial models corresponds to the stationary distribution of an inhomogeneous Markov chain uniquely. Based on this property, we propose an iterative algorithm that is easy to implement and interpret, and is guaranteed to converge. Numerical experiments on synthetic data and real data demonstrate the advantages of our Markov chain based algorithm over existing ones. Our algorithm converges to the MLE with fewer iterations and at a faster convergence rate. The new algorithm is readily applicable to problems such as page ranking or sports ranking data.
****
Towards a Deep and Unified Understanding of Deep Neural Models in NLP
Chaoyu Guan, Xiting Wang, Quanshi Zhang, Runjin Chen, Di He, Xing Xie
We define a unified information-based measure to provide quantitative explanations on how intermediate layers of deep Natural Language Processing (NLP) models leverage information of input words. Our method advances existing explanation methods by addressing issues in coherency and generality. Explanations generated by using our method are consistent and faithful across different timestamps, layers, and models. We show how our method can be applied to four widely used models in NLP and explain their performances on three real-world benchmark datasets.
****
An Investigation of Model-Free Planning
Arthur Guez, Mehdi Mirza, Karol Gregor, Rishabh Kabra, Sebastien Racaniere, Theo

phane Weber, David Raposo, Adam Santoro, Laurent Orseau, Tom Eccles, Greg Wayne,
 David Silver, Timothy Lillicrap
The field of reinforcement learning (RL) is facing increasingly challenging doma
ins with combinatorial complexity. For an RL agent to address these challenges,
it is essential that it can plan effectively. Prior work has typically utilized
an explicit model of the environment, combined with a specific planning algorith
m (such as tree search). More recently, a new family of methods have been propos
ed that learn how to plan, by providing the structure for planning via an induct
ive bias in the function approximator (such as a tree structured neural network)
, trained end-to-end by a model-free RL algorithm. In this paper, we go even fur
ther, and demonstrate empirically that an entirely model-free approach, without
special structure beyond standard neural network components such as convolutiona
l networks and LSTMs, can learn to exhibit many of the characteristics typically
 associated with a model-based planner. We measure our agent’s effectiveness at
planning in terms of its ability to generalize across a combinatorial and irreve
rsible state space, its data efficiency, and its ability to utilize additional t
hinking time. We find that our agent has many of the characteristics that one mi
ght expect to find in a planning algorithm. Furthermore, it exceeds the state-of
-the-art in challenging combinatorial domains such as Sokoban and outperforms ot
her model-free approaches that utilize strong inductive biases toward planning.
****
Humor in Word Embeddings: Cockamamie Gobbledegook for Nincompoops
Limor Gultchin, Genevieve Patterson, Nancy Baym, Nathaniel Swinger, Adam Kalai
While humor is often thought to be beyond the reach of Natural Language Processi
ng, we show that several aspects of single-word humor correlate with simple line
ar directions in Word Embeddings. In particular: (a) the word vectors capture mu
ltiple aspects discussed in humor theories from various disciplines; (b) each in
dividual’s sense of humor can be represented by a vector, which can predict diff
erences in people’s senses of humor on new, unrated, words; and (c) upon cluster
ing humor ratings of multiple demographic groups, different humor preferences em
erge across the different groups. Humor ratings are taken from the work of Engel
thaler and Hills (2017) as well as from an original crowdsourcing study of 120,0
00 words. Our dataset further includes annotations for the theoretically-motivat
ed humor features we identify.
****
Simple Black-box Adversarial Attacks
Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, Kilian Weinberger
We propose an intriguingly simple method for the construction of adversarial ima
ges in the black-box setting. In constrast to the white-box scenario, constructi
ng black-box adversarial images has the additional constraint on query budget, a
nd efficient attacks remain an open problem to date. With only the mild assumpti
on of requiring continuous-valued confidence scores, our highly query-efficient
algorithm utilizes the following simple iterative principle: we randomly sample
a vector from a predefined orthonormal basis and either add or subtract it to th
e target image. Despite its simplicity, the proposed method can be used for both
 untargeted and targeted attacks – resulting in previously unprecedented query e
fficiency in both settings. We demonstrate the efficacy and efficiency of our al
gorithm on several real world settings including the Google Cloud Vision API. We
 argue that our proposed algorithm should serve as a strong baseline for future
black-box attacks, in particular because it is extremely fast and its implementa
tion requires less than 20 lines of PyTorch code.
****
Exploring interpretable LSTM neural networks over multi-variable data
Tian Guo, Tao Lin, Nino Antulov-Fantulin
For recurrent neural networks trained on time series with target and exogenous v
ariables, in addition to accurate prediction, it is also desired to provide inte
rpretable insights into the data. In this paper, we explore the structure of LST
M recurrent neural networks to learn variable-wise hidden states, with the aim t
o capture different dynamics in multi-variable time series and distinguish the c
ontribution of variables to the prediction. With these variable-wise hidden stat

es, a mixture attention mechanism is proposed to model the generative process of the target. Then we develop associated training methods to jointly learn network parameters, variable and temporal importance w.r.t the prediction of the target variable. Extensive experiments on real datasets demonstrate enhanced prediction performance by capturing the dynamics of different variables. Meanwhile, we evaluate the interpretation results both qualitatively and quantitatively. It exhibits the prospect as an end-to-end framework for both forecasting and knowledge extraction over multi-variable data.
****

Learning to Exploit Long-term Relational Dependencies in Knowledge Graphs

Lingbing Guo, Zequn Sun, Wei Hu

We study the problem of knowledge graph (KG) embedding. A widely-established assumption to this problem is that similar entities are likely to have similar relational roles. However, existing related methods derive KG embeddings mainly based on triple-level learning, which lack the capability of capturing long-term relational dependencies of entities. Moreover, triple-level learning is insufficient for the propagation of semantic information among entities, especially for the case of cross-KG embedding. In this paper, we propose recurrent skipping networks (RSNs), which employ a skipping mechanism to bridge the gaps between entities. RSNs integrate recurrent neural networks (RNNs) with residual learning to efficiently capture the long-term relational dependencies within and between KGs. We design an end-to-end framework to support RSNs on different tasks. Our experimental results showed that RSNs outperformed state-of-the-art embedding-based methods for entity alignment and achieved competitive performance for KG completion.
****

Memory-Optimal Direct Convolutions for Maximizing Classification Accuracy in Embedded Applications

Albert Gural, Boris Murmann

In the age of Internet of Things (IoT), embedded devices ranging from ARM Cortex M0s with hundreds of KB of RAM to Arduinos with 2KB RAM are expected to perform increasingly sophisticated classification tasks, such as voice and gesture recognition, activity tracking, and biometric security. While convolutional neural networks (CNNs), together with spectrogram preprocessing, are a natural solution to many of these classification tasks, storage of the network’s activations often exceeds the hard memory constraints of embedded platforms. This paper presents memory-optimal direct convolutions as a way to push classification accuracy as high as possible given strict hardware memory constraints at the expense of extra compute. We therefore explore the opposite end of the compute-memory trade-off curve from standard approaches that minimize latency. We validate the memory-optimal CNN technique with an Arduino implementation of the 10-class MNIST classification task, fitting the network specification, weights, and activations entirely within 2KB SRAM and achieving a state-of-the-art classification accuracy for small-scale embedded systems of 99.15%.
****

IMEXnet A Forward Stable Deep Neural Network

Eldad Haber, Keegan Lensink, Eran Treister, Lars Ruthotto

Deep convolutional neural networks have revolutionized many machine learning and computer vision tasks, however, some remaining key challenges limit their wider use. These challenges include improving the network’s robustness to perturbations of the input image and the limited “field of view” of convolution operators. We introduce the IMEXnet that addresses these challenges by adapting semi-implicit methods for partial differential equations. Compared to similar explicit networks, such as residual networks, our network is more stable, which has recently shown to reduce the sensitivity to small changes in the input features and improve generalization. The addition of an implicit step connects all pixels in each channel of the image and therefore addresses the field of view problem while still being comparable to standard convolutions in terms of the number of parameters and computational complexity. We also present a new dataset for semantic segmentation and demonstrate the effectiveness of our architecture using the NYU Depth dataset.

****
On The Power of Curriculum Learning in Training Deep Networks
Guy Hacohen, Daphna Weinshall
Training neural networks is traditionally done by providing a sequence of random mini-batches sampled uniformly from the entire training data. In this work, we analyze the effect of curriculum learning, which involves the non-uniform sampling of mini-batches, on the training of deep networks, and specifically CNNs trained for image recognition. To employ curriculum learning, the training algorithm must resolve 2 problems: (i) sort the training examples by difficulty; (ii) compute a series of mini-batches that exhibit an increasing level of difficulty. We address challenge (i) using two methods: transfer learning from some competitive "teacher" network, and bootstrapping. In our empirical evaluation, both methods show similar benefits in terms of increased learning speed and improved final performance on test data. We address challenge (ii) by investigating different pacing functions to guide the sampling. The empirical investigation includes a variety of network architectures, using images from CIFAR-10, CIFAR-100 and subsets of ImageNet. We conclude with a novel theoretical analysis of curriculum learning, where we show how it effectively modifies the optimization landscape. We then define the concept of an ideal curriculum, and show that under mild conditions it does not change the corresponding global minimum of the optimization function.
****
Trading Redundancy for Communication: Speeding up Distributed SGD for Non-convex Optimization
Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, Viveck Cadambe
Communication overhead is one of the key challenges that hinders the scalability of distributed optimization algorithms to train large neural networks. In recent years, there has been a great deal of research to alleviate communication cost by compressing the gradient vector or using local updates and periodic model averaging. In this paper, we advocate the use of redundancy towards communication-efficient distributed stochastic algorithms for non-convex optimization. In particular, we, both theoretically and practically, show that by properly infusing redundancy to the training data with model averaging, it is possible to significantly reduce the number of communication rounds. To be more precise, we show that redundancy reduces residual error in local averaging, thereby reaching the same level of accuracy with fewer rounds of communication as compared with previous algorithms. Empirical studies on CIFAR10, CIFAR100 and ImageNet datasets in a distributed environment complement our theoretical results; they show that our algorithms have additional beneficial aspects including tolerance to failures, as well as greater gradient diversity.
****
Learning Latent Dynamics for Planning from Pixels
Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, James Davidson
Planning has been very successful for control tasks with known environment dynamics. To leverage planning in unknown environments, the agent needs to learn the dynamics from interactions with the world. However, learning dynamics models that are accurate enough for planning has been a long-standing challenge, especially in image-based domains. We propose the Deep Planning Network (PlaNet), a purely model-based agent that learns the environment dynamics from images and chooses actions through fast online planning in latent space. To achieve high performance, the dynamics model must accurately predict the rewards ahead for multiple time steps. We approach this using a latent dynamics model with both deterministic and stochastic transition components. Moreover, we propose a multi-step variational inference objective that we name latent overshooting. Using only pixel observations, our agent solves continuous control tasks with contact dynamics, partial observability, and sparse rewards, which exceed the difficulty of tasks that were previously solved by planning with learned models. PlaNet uses substantially fewer episodes and reaches final performance close to and sometimes higher than strong model-free algorithms.

****
Neural Separation of Observed and Unobserved Distributions
Tavi Halperin, Ariel Ephrat, Yedid Hoshen
Separating mixed distributions is a long standing challenge for machine learning
 and signal processing. Most current methods either rely on making strong assump
tions on the source distributions or rely on having training samples of each sou
rce in the mixture. In this work, we introduce a new method—Neural Egg Separatio
n—to tackle the scenario of extracting a signal from an unobserved distribution
additively mixed with a signal from an observed distribution. Our method iterati
vely learns to separate the known distribution from progressively finer estimate
s of the unknown distribution. In some settings, Neural Egg Separation is initia
lization sensitive, we therefore introduce Latent Mixture Masking which ensures
a good initialization. Extensive experiments on audio and image separation tasks
 show that our method outperforms current methods that use the same level of sup
ervision, and often achieves similar performance to full supervision.
****
Grid-Wise Control for Multi-Agent Reinforcement Learning in Video Game AI
Lei Han, Peng Sun, Yali Du, Jiechao Xiong, Qing Wang, Xinghai Sun, Han Liu, Tong
 Zhang
We consider the problem of multi-agent reinforcement learning (MARL) in video ga
me AI, where the agents are located in a spatial grid-world environment and the
number of agents varies both within and across episodes. The challenge is to fle
xibly control an arbitrary number of agents while achieving effective collaborat
ion. Existing MARL methods usually suffer from the trade-off between these two c
onsiderations. To address the issue, we propose a novel architecture that learns
 a spatial joint representation of all the agents and outputs grid-wise actions.
 Each agent will be controlled independently by taking the action from the grid
it occupies. By viewing the state information as a grid feature map, we employ a
 convolutional encoder-decoder as the policy network. This architecture naturall
y promotes agent communication because of the large receptive field provided by
the stacked convolutional layers. Moreover, the spatially shared convolutional p
arameters enable fast parallel exploration that the experiences discovered by on
e agent can be immediately transferred to others. The proposed method can be con
veniently integrated with general reinforcement learning algorithms, e.g., PPO a
nd Q-learning. We demonstrate the effectiveness of the proposed method in extens
ive challenging multi-agent tasks in StarCraft II.
****
Dimension-Wise Importance Sampling Weight Clipping for Sample-Efficient Reinforc
ement Learning
Seungyul Han, Youngchul Sung
In importance sampling (IS)-based reinforcement learning algorithms such as Prox
imal Policy Optimization (PPO), IS weights are typically clipped to avoid large
variance in learning. However, policy update from clipped statistics induces lar
ge bias in tasks with high action dimensions, and bias from clipping makes it di
fficult to reuse old samples with large IS weights. In this paper, we consider P
PO, a representative on-policy algorithm, and propose its improvement by dimensi
on-wise IS weight clipping which separately clips the IS weight of each action d
imension to avoid large bias and adaptively controls the IS weight to bound poli
cy update from the current policy. This new technique enables efficient learning
 for high action-dimensional tasks and reusing of old samples like in off-policy
 learning to increase the sample efficiency. Numerical results show that the pro
posed new algorithm outperforms PPO and other RL algorithms in various Open AI G
ym tasks.
****
Complexity of Linear Regions in Deep Networks
Boris Hanin, David Rolnick
It is well-known that the expressivity of a neural network depends on its archit
ecture, with deeper networks expressing more complex functions. In the case of n
etworks that compute piecewise linear functions, such as those with ReLU activat
ion, the number of distinct linear regions is a natural measure of expressivity.

It is possible to construct networks with merely a single region, or for which the number of linear regions grows exponentially with depth; it is not clear where within this range most networks fall in practice, either before or after training. In this paper, we provide a mathematical framework to count the number of linear regions of a piecewise linear network and measure the volume of the boundaries between these regions. In particular, we prove that for networks at initialization, the average number of regions along any one-dimensional subspace grows linearly in the total number of neurons, far below the exponential upper bound. We also find that the average distance to the nearest region boundary at initialization scales like the inverse of the number of neurons. Our theory suggests that, even after training, the number of linear regions is far below exponential, an intuition that matches our empirical observations. We conclude that the practical expressivity of neural networks is likely far below that of the theoretical maximum, and that this gap can be quantified.

****

Importance Sampling Policy Evaluation with an Estimated Behavior Policy

Josiah Hanna, Scott Niekum, Peter Stone

We consider the problem of off-policy evaluation in Markov decision processes. Off-policy evaluation is the task of evaluating the expected return of one policy with data generated by a different, behavior policy. Importance sampling is a technique for off-policy evaluation that re-weights off-policy returns to account for differences in the likelihood of the returns between the two policies. In this paper, we study importance sampling with an estimated behavior policy where the behavior policy estimate comes from the same set of data used to compute the importance sampling estimate. We find that this estimator often lowers the mean squared error of off-policy evaluation compared to importance sampling with the true behavior policy or using a behavior policy that is estimated from a separate data set. Intuitively, estimating the behavior policy in this way corrects for error due to sampling in the action-space. Our empirical results also extend to other popular variants of importance sampling and show that estimating a non-Markovian behavior policy can further lower large-sample mean squared error even when the true behavior policy is Markovian.

****

Doubly-Competitive Distribution Estimation

Yi Hao, Alon Orlitsky

Distribution estimation is a statistical-learning cornerstone. Its classical min-max formulation minimizes the estimation error for the worst distribution, hence under-performs for practical distributions that, like power-law, are often rather simple. Modern research has therefore focused on two frameworks: structural estimation that improves learning accuracy by assuming a simple structure of the underlying distribution; and competitive, or instance-optimal, estimation that achieves the performance of a genie aided estimator up to a small excess error that vanishes as the sample size grows, regardless of the distribution. This paper combines and strengthens the two frameworks. It designs a single estimator whose excess error vanishes both at a universal rate as the sample size grows, as well as when the (unknown) distribution gets simpler. We show that the resulting algorithm significantly improves the performance guarantees for numerous competitive- and structural-estimation results. The algorithm runs in near-linear time and is robust to model misspecification and domain-symbol permutations.

****

Random Shuffling Beats SGD after Finite Epochs

Jeff Haochen, Suvrit Sra

A long-standing problem in stochastic optimization is proving that \rsgd, the without-replacement version of \sgd, converges faster than the usual with-replacement \sgd. Building upon \citep{gurbuzbalaban2015random}, we present the first (to our knowledge) non-asymptotic results for this problem by proving that after a reasonable number of epochs \rsgd converges faster than \sgd. Specifically, we prove that for strongly convex, second-order smooth functions, the iterates of \rsgd converge to the optimal solution as $\mathcal{O}(\nicefrac{1}{T^2} + \nicefrac{n^3}{T^3})$, where n is the number of components in the objective, and T

is number of iterations. This result implies that after $\mathcal{O}(\sqrt{n})$ epochs, \rsgd is strictly better than \sgd (which converges as $\mathcal{O}(\nicefrac{1}{T})$). The key step toward showing this better dependence on T is the introduction of n into the bound; and as our analysis shows, in general a dependence on n is unavoidable without further changes. To understand how \rsgd works in practice, we further explore two empirically useful settings: data sparsity and over-parameterization. For sparse data, \rsgd has the rate $\mathcal{O}\left(\frac{1}{T^2}\right)$, again strictly better than \sgd. Under a setting closely related to over-parameterization, \rsgd is shown to converge faster than \sgd after any arbitrary number of iterations. Finally, we extend the analysis of \rsgd to smooth non-convex and convex functions.
****

Submodular Maximization beyond Non-negativity: Guarantees, Fast Algorithms, and Applications
Chris Harshaw, Moran Feldman, Justin Ward, Amin Karbasi
It is generally believed that submodular functions–and the more general class of γ-weakly submodular functions–may only be optimized under the non-negativity assumption $f(S) \geq 0$. In this paper, we show that once the function is expressed as the difference $f = g - c$, where g is monotone, non-negative, and γ-weakly submodular and c is non-negative modular, then strong approximation guarantees may be obtained. We present an algorithm for maximizing $g - c$ under a k-cardinality constraint which produces a random feasible set S such that $\mathbb{E}[g(S) -c(S)] \geq (1 - e^{-\gamma} - \epsilon) g(\opt) - c(\opt)$, whose running time is $O (\frac{n}{\epsilon} \log^2 \frac{1}{\epsilon})$, independent of k. We extend these results to the unconstrained setting by describing an algorithm with the same approximation guarantees and faster $O(n \frac{1}{\epsilon} \log\frac{1}{\epsilon})$ runtime. The main techniques underlying our algorithms are two-fold: the use of a surrogate objective which varies the relative importance between g and c throughout the algorithm, and a geometric sweep over possible γ values. Our algorithmic guarantees are complemented by a hardness result showing that no polynomial-time algorithm which accesses g through a value oracle can do better. We empirically demonstrate the success of our algorithms by applying them to experimental design on the Boston Housing dataset and directed vertex cover on the Email EU dataset.
****

Per-Decision Option Discounting
Anna Harutyunyan, Peter Vrancx, Philippe Hamel, Ann Nowe, Doina Precup
In order to solve complex problems an agent must be able to reason over a sufficiently long horizon. Temporal abstraction, commonly modeled through options, offers the ability to reason at many timescales, but the horizon length is still determined by the discount factor of the underlying Markov Decision Process. We propose a modification to the options framework that naturally scales the agent’s horizon with option length. We show that the proposed option-step discount controls a bias-variance trade-off, with larger discounts (counter-intuitively) leading to less estimation variance.
****

Submodular Observation Selection and Information Gathering for Quadratic Models
Abolfazl Hashemi, Mahsa Ghasemi, Haris Vikalo, Ufuk Topcu
We study the problem of selecting most informative subset of a large observation set to enable accurate estimation of unknown parameters. This problem arises in a variety of settings in machine learning and signal processing including feature selection, phase retrieval, and target localization. Since for quadratic measurement models the moment matrix of the optimal estimator is generally unknown, majority of prior work resorts to approximation techniques such as linearization of the observation model to optimize the alphabetical optimality criteria of an approximate moment matrix. Conversely, by exploiting a connection to the classical Van Trees’ inequality, we derive new alphabetical optimality criteria without distorting the relational structure of the observation model. We further show that under certain conditions on parameters of the problem these optimality criteria are monotone and (weak) submodular set functions. These results enable us t

o develop an efficient greedy observation selection algorithm uniquely tailored
for quadratic models, and provide theoretical bounds on its achievable utility.
****
Understanding and Controlling Memory in Recurrent Neural Networks
Doron Haviv, Alexander Rivkind, Omri Barak
To be effective in sequential data processing, Recurrent Neural Networks (RNNs)
are required to keep track of past events by creating memories. While the relati
on between memories and the network’s hidden state dynamics was established over
 the last decade, previous works in this direction were of a predominantly descr
iptive nature focusing mainly on locating the dynamical objects of interest. In
particular, it remained unclear how dynamical observables affect the performance
, how they form and whether they can be manipulated. Here, we utilize different
training protocols, datasets and architectures to obtain a range of networks sol
ving a delayed classification task with similar performance, alongside substanti
al differences in their ability to extrapolate for longer delays. We analyze the
 dynamics of the network’s hidden state, and uncover the reasons for this differ
ence. Each memory is found to be associated with a nearly steady state of the dy
namics which we refer to as a ’slow point’. Slow point speeds predict extrapolat
ion performance across all datasets, protocols and architectures tested. Further
more, by tracking the formation of the slow points we are able to understand the
 origin of differences between training protocols. Finally, we propose a novel r
egularization technique that is based on the relation between hidden state speed
s and memory longevity. Our technique manipulates these speeds, thereby leading
to a dramatic improvement in memory robustness over time, and could pave the way
 for a new class of regularization methods.
****
On the Impact of the Activation function on Deep Neural Networks Training
Soufiane Hayou, Arnaud Doucet, Judith Rousseau
The weight initialization and the activation function of deep neural networks ha
ve a crucial impact on the performance of the training procedure. An inappropria
te selection can lead to the loss of information of the input during forward pro
pagation and the exponential vanishing/exploding of gradients during back-propag
ation. Understanding the theoretical properties of untrained random networks is
key to identifying which deep networks may be trained successfully as recently d
emonstrated by Samuel et al. (2017) who showed that for deep feedforward neural
networks only a specific choice of hyperparameters known as the ‘Edge of Chaos’
can lead to good performance. While the work by Samuel et al. (2017) discuss tra
inability issues, we focus here on training acceleration and overall performance
. We give a comprehensive theoretical analysis of the Edge of Chaos and show tha
t we can indeed tune the initialization parameters and the activation function i
n order to accelerate the training and improve the performance.
****
Provably Efficient Maximum Entropy Exploration
Elad Hazan, Sham Kakade, Karan Singh, Abby Van Soest
Suppose an agent is in a (possibly unknown) Markov Decision Process in the absen
ce of a reward signal, what might we hope that an agent can efficiently learn to
 do? This work studies a broad class of objectives that are defined solely as fu
nctions of the state-visitation frequencies that are induced by how the agent be
haves. For example, one natural, intrinsically defined, objective problem is for
 the agent to learn a policy which induces a distribution over state space that
is as uniform as possible, which can be measured in an entropic sense. We provid
e an efficient algorithm to optimize such such intrinsically defined objectives,
 when given access to a black box planning oracle (which is robust to function a
pproximation). Furthermore, when restricted to the tabular setting where we have
 sample based access to the MDP, our proposed algorithm is provably efficient, b
oth in terms of its sample and computational complexities. Key to our algorithmi
c methodology is utilizing the conditional gradient method (a.k.a. the Frank-Wol
fe algorithm) which utilizes an approximate MDP solver.
****
On the Long-term Impact of Algorithmic Decision Policies: Effort Unfairness and

Feature Segregation through Social Learning
Hoda Heidari, Vedant Nanda, Krishna Gummadi
Most existing notions of algorithmic fairness are one-shot: they ensure some form of allocative equality at the time of decision making, but do not account for the adverse impact of the algorithmic decisions today on the long-term welfare and prosperity of certain segments of the population. We take a broader perspective on algorithmic fairness. We propose an effort-based measure of fairness and present a data-driven framework for characterizing the long-term impact of algorithmic policies on reshaping the underlying population. Motivated by the psychological literature on social learning and the economic literature on equality of opportunity, we propose a micro-scale model of how individuals may respond to decision-making algorithms. We employ existing measures of segregation from sociology and economics to quantify the resulting macro- scale population-level change. Importantly, we observe that different models may shift the group- conditional distribution of qualifications in different directions. Our findings raise a number of important questions regarding the formalization of fairness for decision-making models.
****
Graph Resistance and Learning from Pairwise Comparisons
Julien Hendrickx, Alexander Olshevsky, Venkatesh Saligrama
We consider the problem of learning the qualities of a collection of items by performing noisy comparisons among them. Following the standard paradigm, we assume there is a fixed “comparison graph” and every neighboring pair of items in this graph is compared k times according to the Bradley-Terry-Luce model (where the probability than an item wins a comparison is proportional the item quality). We are interested in how the relative error in quality estimation scales with the comparison graph in the regime where k is large. We show that, asymptotically, the relevant graph-theoretic quantity is the square root of the resistance of the comparison graph. Specifically, we provide an algorithm with relative error decay that scales with the square root of the graph resistance, and provide a lower bound showing that (up to log factors) a better scaling is impossible. The performance guarantee of our algorithm, both in terms of the graph and the skewness of the item quality distribution, significantly outperforms earlier results.
****
Using Pre-Training Can Improve Model Robustness and Uncertainty
Dan Hendrycks, Kimin Lee, Mantas Mazeika
He et al. (2018) have called into question the utility of pre-training by showing that training from scratch can often yield similar performance to pre-training. We show that although pre-training may not improve performance on traditional classification metrics, it improves model robustness and uncertainty estimates. Through extensive experiments on label corruption, class imbalance, adversarial examples, out-of-distribution detection, and confidence calibration, we demonstrate large gains from pre-training and complementary effects with task-specific methods. We show approximately a 10% absolute improvement over the previous state-of-the-art in adversarial robustness. In some cases, using pre-training without task-specific methods also surpasses the state-of-the-art, highlighting the need for pre-training when evaluating future methods on robustness and uncertainty tasks.
****
Flow++: Improving Flow-Based Generative Models with Variational Dequantization and Architecture Design
Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, Pieter Abbeel
Flow-based generative models are powerful exact likelihood models with efficient sampling and inference. Despite their computational efficiency, flow-based models generally have much worse density modeling performance compared to state-of-the-art autoregressive models. In this paper, we investigate and improve upon three limiting design choices employed by flow-based models in prior work: the use of uniform noise for dequantization, the use of inexpressive affine flows, and the use of purely convolutional conditioning networks in coupling layers. Based on our findings, we propose Flow++, a new flow-based model that is now the state-

of-the-art non-autoregressive model for unconditional density estimation on standard image benchmarks. Our work has begun to close the significant performance gap that has so far existed between autoregressive models and flow-based models.
****
Population Based Augmentation: Efficient Learning of Augmentation Policy Schedules
Daniel Ho, Eric Liang, Xi Chen, Ion Stoica, Pieter Abbeel
A key challenge in leveraging data augmentation for neural network training is choosing an effective augmentation policy from a large search space of candidate operations. Properly chosen augmentation policies can lead to significant generalization improvements; however, state-of-the-art approaches such as AutoAugment are computationally infeasible to run for the ordinary user. In this paper, we introduce a new data augmentation algorithm, Population Based Augmentation (PBA), which generates nonstationary augmentation policy schedules instead of a fixed augmentation policy. We show that PBA can match the performance of AutoAugment on CIFAR-10, CIFAR-100, and SVHN, with three orders of magnitude less overall compute. On CIFAR-10 we achieve a mean test error of 1.46%, which is a slight improvement upon the current state-of-the-art. The code for PBA is open source and is available at https://github.com/arcelien/pba.
****
Collective Model Fusion for Multiple Black-Box Experts
Minh Hoang, Nghia Hoang, Bryan Kian Hsiang Low, Carleton Kingsford
Model fusion is a fundamental problem in collec-tive machine learning (ML) where independentexperts with heterogeneous learning architecturesare required to combine expertise to improve pre-dictive performance. This is particularly chal-lenging in information-sensitive domains whereexperts do not have access to each other's internalarchitecture and local data. This paper presentsthe first collective model fusion framework formultiple experts with heterogeneous black-box ar-chitectures. The proposed method will enable thisby addressing the key issues of how black-boxexperts interact to understand the predictive be-haviors of one another; how these understandingscan be represented and shared efficiently amongthemselves; and how the shared understandingscan be combined to generate high-quality consen-sus prediction. The performance of the resultingframework is analyzed theoretically and demon-strated empirically on several datasets.
****
Connectivity-Optimized Representation Learning via Persistent Homology
Christoph Hofer, Roland Kwitt, Marc Niethammer, Mandar Dixit
We study the problem of learning representations with controllable connectivity properties. This is beneficial in situations when the imposed structure can be leveraged upstream. In particular, we control the connectivity of an autoencoder's latent space via a novel type of loss, operating on information from persistent homology. Under mild conditions, this loss is differentiable and we present a theoretical analysis of the properties induced by the loss. We choose one-class learning as our upstream task and demonstrate that the imposed structure enables informed parameter selection for modeling the in-class distribution via kernel density estimators. Evaluated on computer vision data, these one-class models exhibit competitive performance and, in a low sample size regime, outperform other methods by a large margin. Notably, our results indicate that a single autoencoder, trained on auxiliary (unlabeled) data, yields a mapping into latent space that can be reused across datasets for one-class learning.
****
Better generalization with less data using robust gradient descent
Matthew Holland, Kazushi Ikeda
For learning tasks where the data (or losses) may be heavy-tailed, algorithms based on empirical risk minimization may require a substantial number of observations in order to perform well off-sample. In pursuit of stronger performance under weaker assumptions, we propose a technique which uses a cheap and robust iterative estimate of the risk gradient, which can be easily fed into any steepest descent procedure. Finite-sample risk bounds are provided under weak moment assumptions on the loss gradient. The algorithm is simple to implement, and empirical

tests using simulations and real-world data illustrate that more efficient and reliable learning is possible without prior knowledge of the loss tails.
****
Emerging Convolutions for Generative Normalizing Flows
Emiel Hoogeboom, Rianne Van Den Berg, Max Welling
Generative flows are attractive because they admit exact likelihood optimization and efficient image synthesis. Recently, Kingma & Dhariwal (2018) demonstrated with Glow that generative flows are capable of generating high quality images. We generalize the 1 {\texttimes} 1 convolutions proposed in Glow to invertible d {\texttimes} d convolutions, which are more flexible since they operate on both channel and spatial axes. We propose two methods to produce invertible convolutions, that have receptive fields identical to standard convolutions: Emerging convolutions are obtained by chaining specific autoregressive convolutions, and periodic convolutions are decoupled in the frequency domain. Our experiments show that the flexibility of d {\texttimes} d convolutions significantly improves the performance of generative flow models on galaxy images, CIFAR10 and ImageNet.
****
Nonconvex Variance Reduced Optimization with Arbitrary Sampling
Samuel Horváth, Peter Richtarik
We provide the first importance sampling variants of variance reduced algorithms for empirical risk minimization with non-convex loss functions. In particular, we analyze non-convex versions of \texttt{SVRG}, \texttt{SAGA} and \texttt{SARAH}. Our methods have the capacity to speed up the training process by an order of magnitude compared to the state of the art on real datasets. Moreover, we also improve upon current mini-batch analysis of these methods by proposing importance sampling for minibatches in this setting. Surprisingly, our approach can in some regimes lead to superlinear speedup with respect to the minibatch size, which is not usually present in stochastic optimization. All the above results follow from a general analysis of the methods which works with arbitrary sampling, i.e., fully general randomized strategy for the selection of subsets of examples to be sampled in each iteration. Finally, we also perform a novel importance sampling analysis of \texttt{SARAH} in the convex setting.
****
Parameter-Efficient Transfer Learning for NLP
Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, Sylvain Gelly
Fine-tuning large pretrained models is an effective transfer mechanism in NLP. However, in the presence of many downstream tasks, fine-tuning is parameter inefficient: an entire new model is required for every task. As an alternative, we propose transfer with adapter modules. Adapter modules yield a compact and extensible model; they add only a few trainable parameters per task, and new tasks can be added without revisiting previous ones. The parameters of the original network remain fixed, yielding a high degree of parameter sharing. To demonstrate adapter’s effectiveness, we transfer the recently proposed BERT Transformer model to 26 diverse text classification tasks, including the GLUE benchmark. Adapters attain near state-of-the-art performance, whilst adding only a few parameters per task. On GLUE, we attain within $0.8%$ of the performance of full fine-tuning, adding only $3.6%$ parameters per task. By contrast, fine-tuning trains $100%$ of the parameters per task.
****
Stay With Me: Lifetime Maximization Through Heteroscedastic Linear Bandits With Reneging
Ping-Chun Hsieh, Xi Liu, Anirban Bhattacharya, P R Kumar
Sequential decision making for lifetime maximization is a critical problem in many real-world applications, such as medical treatment and portfolio selection. In these applications, a “reneging” phenomenon, where participants may disengage from future interactions after observing an unsatisfiable outcome, is rather prevalent. To address the above issue, this paper proposes a model of heteroscedastic linear bandits with reneging, which allows each participant to have a distinct “satisfaction level," with any interaction outcome falling short of that level

resulting in that participant reneging. Moreover, it allows the variance of the outcome to be context-dependent. Based on this model, we develop a UCB-type policy, namely HR-UCB, and prove that it achieves $\mathcal{O}\big(\sqrt{{T}(\log({T}))^{3}}\big)$ regret. Finally, we validate the performance of HR-UCB via simulations.
****
Finding Mixed Nash Equilibria of Generative Adversarial Networks
Ya-Ping Hsieh, Chen Liu, Volkan Cevher
Generative adversarial networks (GANs) are known to achieve the state-of-the-art performance on various generative tasks, but these results come at the expense of a notoriously difficult training phase. Current training strategies typically draw a connection to optimization theory, whose scope is restricted to local convergence due to the presence of non-convexity. In this work, we tackle the training of GANs by rethinking the problem formulation from the mixed Nash Equilibria (NE) perspective. Via a classical lifting trick, we show that essentially all existing GAN objectives can be relaxed into their mixed strategy forms, whose global optima can be solved via sampling, in contrast to the exclusive use of optimization framework in previous work. We further propose a mean-approximation sampling scheme, which allows to systematically exploit methods for bi-affine games to delineate novel, practical training algorithms of GANs. Finally, we provide experimental evidence that our approach yields comparable or superior results to contemporary training algorithms, and outperforms classical methods such as SGD, Adam, and RMSProp.
****
Classification from Positive, Unlabeled and Biased Negative Data
Yu-Guan Hsieh, Gang Niu, Masashi Sugiyama
In binary classification, there are situations where negative (N) data are too diverse to be fully labeled and we often resort to positive-unlabeled (PU) learning in these scenarios. However, collecting a non-representative N set that contains only a small portion of all possible N data can often be much easier in practice. This paper studies a novel classification framework which incorporates such biased N (bN) data in PU learning. We provide a method based on empirical risk minimization to address this PUbN classification problem. Our approach can be regarded as a novel example-weighting algorithm, with the weight of each example computed through a preliminary step that draws inspiration from PU learning. We also derive an estimation error bound for the proposed method. Experimental results demonstrate the effectiveness of our algorithm in not only PUbN learning scenarios but also ordinary PU learning scenarios on several benchmark datasets.
****
Bayesian Deconditional Kernel Mean Embeddings
Kelvin Hsu, Fabio Ramos
Conditional kernel mean embeddings form an attractive nonparametric framework for representing conditional means of functions, describing the observation processes for many complex models. However, the recovery of the original underlying function of interest whose conditional mean was observed is a challenging inference task. We formalize deconditional kernel mean embeddings as a solution to this inverse problem, and show that it can be naturally viewed as a nonparametric Bayes' rule. Critically, we introduce the notion of task transformed Gaussian processes and establish deconditional kernel means as their posterior predictive mean. This connection provides Bayesian interpretations and uncertainty estimates for deconditional kernel mean embeddings, explains their regularization hyperparameters, and reveals a marginal likelihood for kernel hyperparameter learning. These revelations further enable practical applications such as likelihood-free inference and learning sparse representations for big data.
****
Faster Stochastic Alternating Direction Method of Multipliers for Nonconvex Optimization
Feihu Huang, Songcan Chen, Heng Huang
In this paper, we propose a faster stochastic alternating direction method of multipliers (ADMM) for nonconvex optimization by using a new stochastic path-integ

rated differential estimator (SPIDER), called as SPIDER-ADMM. Moreover, we prove that the SPIDER-ADMM achieves a record-breaking incremental first-order oracle (IFO) complexity for finding an ϵ-approximate solution. As one of major contribution of this paper, we provide a new theoretical analysis framework for nonconvex stochastic ADMM methods with providing the optimal IFO complexity. Based on this new analysis framework, we study the unsolved optimal IFO complexity of the existing non-convex SVRG-ADMM and SAGA-ADMM methods, and prove their the optimal IFO complexity. Thus, the SPIDER-ADMM improves the existing stochastic ADMM methods. Moreover, we extend SPIDER-ADMM to the online setting, and propose a faster online SPIDER-ADMM. Our theoretical analysis also derives the IFO complexity of the online SPIDER-ADMM. Finally, the experimental results on benchmark datasets validate that the proposed algorithms have faster convergence rate than the existing ADMM algorithms for nonconvex optimization.
****
Unsupervised Deep Learning by Neighbourhood Discovery
Jiabo Huang, Qi Dong, Shaogang Gong, Xiatian Zhu
Deep convolutional neural networks (CNNs) have demonstrated remarkable success in computer vision by supervisedly learning strong visual feature representations. However, training CNNs relies heavily on the availability of exhaustive training data annotations, limiting significantly their deployment and scalability in many application scenarios. In this work, we introduce a generic unsupervised deep learning approach to training deep models without the need for any manual label supervision. Specifically, we progressively discover sample anchored/centred neighbourhoods to reason and learn the underlying class decision boundaries iteratively and accumulatively. Every single neighbourhood is specially formulated so that all the member samples can share the same unseen class labels at high probability for facilitating the extraction of class discriminative feature representations during training. Experiments on image classification show the performance advantages of the proposed method over the state-of-the-art unsupervised learning models on six benchmarks including both coarse-grained and fine-grained object image categorisation.
****
Detecting Overlapping and Correlated Communities without Pure Nodes: Identifiability and Algorithm
Kejun Huang, Xiao Fu
Many machine learning problems come in the form of networks with relational data between entities, and one of the key unsupervised learning tasks is to detect communities in such a network. We adopt the mixed-membership stochastic blockmodel as the underlying probabilistic model, and give conditions under which the memberships of a subset of nodes can be uniquely identified. Our method starts by constructing a second-order graph moment, which can be shown to converge to a specific product of the true parameters as the size of the network increases. To correctly recover the true membership parameters, we formulate an optimization problem using insights from convex geometry. We show that if the true memberships satisfy a so-called sufficiently scattered condition, then solving the proposed problem correctly identifies the ground truth. We also propose an efficient algorithm for detecting communities, which is significantly faster than prior work and with better convergence properties. Experiments on synthetic and real data justify the validity of the proposed learning framework for network data.
****
Hierarchical Importance Weighted Autoencoders
Chin-Wei Huang, Kris Sankaran, Eeshan Dhekane, Alexandre Lacoste, Aaron Courville
Importance weighted variational inference (Burda et al., 2015) uses multiple i.i.d. samples to have a tighter variational lower bound. We believe a joint proposal has the potential of reducing the number of redundant samples, and introduce a hierarchical structure to induce correlation. The hope is that the proposals would coordinate to make up for the error made by one another to reduce the variance of the importance estimator. Theoretically, we analyze the condition under which convergence of the estimator variance can be connected to convergence of th

e lower bound. Empirically, we confirm that maximization of the lower bound does implicitly minimize variance. Further analysis shows that this is a result of negative correlation induced by the proposed hierarchical meta sampling scheme, and performance of inference also improves when the number of samples increases.
****
Stable and Fair Classification
Lingxiao Huang, Nisheeth Vishnoi
In a recent study, Friedler et al. pointed out that several fair classification algorithms are not stable with respect to variations in the training set – a crucial consideration in several applications. Motivated by their work, we study the problem of designing classification algorithms that are both fair and stable. We propose an extended framework based on fair classification algorithms that are formulated as optimization problems, by introducing a stability-focused regularization term. Theoretically, we prove an additional stability guarantee, that was lacking in fair classification algorithms, and also provide an accuracy guarantee for our extended framework. Our accuracy guarantee can be used to inform the selection of the regularization parameter in our framework. We assess the benefits of our approach empirically by extending several fair classification algorithms that are shown to achieve the best balance between fairness and accuracy over the \textbf{Adult} dataset. Our empirical results show that our extended framework indeed improves the stability at only a slight sacrifice in accuracy.
****
Addressing the Loss-Metric Mismatch with Adaptive Loss Alignment
Chen Huang, Shuangfei Zhai, Walter Talbott, Miguel Bautista Martin, Shih-Yu Sun, Carlos Guestrin, Josh Susskind
In most machine learning training paradigms a fixed, often handcrafted, loss function is assumed to be a good proxy for an underlying evaluation metric. In this work we assess this assumption by meta-learning an adaptive loss function to directly optimize the evaluation metric. We propose a sample efficient reinforcement learning approach for adapting the loss dynamically during training. We empirically show how this formulation improves performance by simultaneously optimizing the evaluation metric and smoothing the loss landscape. We verify our method in metric learning and classification scenarios, showing considerable improvements over the state-of-the-art on a diverse set of tasks. Importantly, our method is applicable to a wide range of loss functions and evaluation metrics. Furthermore, the learned policies are transferable across tasks and data, demonstrating the versatility of the method.
****
Causal Discovery and Forecasting in Nonstationary Environments with State-Space Models
Biwei Huang, Kun Zhang, Mingming Gong, Clark Glymour
In many scientific fields, such as economics and neuroscience, we are often faced with nonstationary time series, and concerned with both finding causal relations and forecasting the values of variables of interest, both of which are particularly challenging in such nonstationary environments. In this paper, we study causal discovery and forecasting for nonstationary time series. By exploiting a particular type of state-space model to represent the processes, we show that nonstationarity helps to identify the causal structure, and that forecasting naturally benefits from learned causal knowledge. Specifically, we allow changes in both causal strengths and noise variances in the nonlinear state-space models, which, interestingly, renders both the causal structure and model parameters identifiable. Given the causal model, we treat forecasting as a problem in Bayesian inference in the causal model, which exploits the time-varying property of the data and adapts to new observations in a principled manner. Experimental results on synthetic and real-world data sets demonstrate the efficacy of the proposed methods.
****
Composing Entropic Policies using Divergence Correction
Jonathan Hunt, Andre Barreto, Timothy Lillicrap, Nicolas Heess
Composing skills mastered in one task to solve novel tasks promises dramatic imp

rovements in the data efficiency of reinforcement learning. Here, we analyze two recent works composing behaviors represented in the form of action-value functions and show that they perform poorly in some situations. As part of this analysis, we extend an important generalization of policy improvement to the maximum entropy framework and introduce an algorithm for the practical implementation of successor features in continuous action spaces. Then we propose a novel approach which addresses the failure cases of prior work and, in principle, recovers the optimal policy during transfer. This method works by explicitly learning the (discounted, future) divergence between base policies. We study this approach in the tabular case and on non-trivial continuous control problems with compositional structure and show that it outperforms or matches existing methods across all tasks considered.
****
HexaGAN: Generative Adversarial Nets for Real World Classification
Uiwon Hwang, Dahuin Jung, Sungroh Yoon
Most deep learning classification studies assume clean data. However, when dealing with the real world data, we encounter three problems such as 1) missing data, 2) class imbalance, and 3) missing label problems. These problems undermine the performance of a classifier. Various preprocessing techniques have been proposed to mitigate one of these problems, but an algorithm that assumes and resolves all three problems together has not been proposed yet. In this paper, we propose HexaGAN, a generative adversarial network framework that shows promising classification performance for all three problems. We interpret the three problems from a single perspective to solve them jointly. To enable this, the framework consists of six components, which interact with each other. We also devise novel loss functions corresponding to the architecture. The designed loss functions allow us to achieve state-of-the-art imputation performance, with up to a 14% improvement, and to generate high-quality class-conditional data. We evaluate the classification performance (F1-score) of the proposed method with 20% missingness and confirm up to a 5% improvement in comparison with the performance of combinations of state-of-the-art methods.
****
Overcoming Mean-Field Approximations in Recurrent Gaussian Process Models
Alessandro Davide Ialongo, Mark Van Der Wilk, James Hensman, Carl Edward Rasmussen
We identify a new variational inference scheme for dynamical systems whose transition function is modelled by a Gaussian process. Inference in this setting has either employed computationally intensive MCMC methods, or relied on factorisations of the variational posterior. As we demonstrate in our experiments, the factorisation between latent system states and transition function can lead to a miscalibrated posterior and to learning unnecessarily large noise terms. We eliminate this factorisation by explicitly modelling the dependence between state trajectories and the low-rank representation of our Gaussian process posterior. Samples of the latent states can then be tractably generated by conditioning on this representation. The method we obtain gives better predictive performance and more calibrated estimates of the transition function, yet maintains the same time and space complexities as mean-field methods.
****
Learning Structured Decision Problems with Unawareness
Craig Innes, Alex Lascarides
Structured models of decision making often assume an agent is aware of all possible states and actions in advance. This assumption is sometimes untenable. In this paper, we learn Bayesian Decision Networks from both domain exploration and expert assertions in a way which guarantees convergence to optimal behaviour, even when the agent starts unaware of actions or belief variables that are critical to success. Our experiments show that our agent learns optimal behaviour on both small and large decision problems, and that allowing an agent to conserve information upon making new discoveries results in faster convergence.
****
Phase transition in PCA with missing data: Reduced signal-to-noise ratio, not sa

mple size!
Niels Ipsen, Lars Kai Hansen
How does missing data affect our ability to learn signal structures? It has been shown that learning signal structure in terms of principal components is dependent on the ratio of sample size and dimensionality and that a critical number of observations is needed before learning starts (Biehl and Mietzner, 1993). Here we generalize this analysis to include missing data. Probabilistic principal component analysis is regularly used for estimating signal structures in datasets with missing data. Our analytic result suggest that the effect of missing data is to effectively reduce signal-to-noise ratio rather than - as generally believed - to reduce sample size. The theory predicts a phase transition in the learning curves and this is indeed found both in simulation data and in real datasets.
****
Actor-Attention-Critic for Multi-Agent Reinforcement Learning
Shariq Iqbal, Fei Sha
Reinforcement learning in multi-agent scenarios is important for real-world applications but presents challenges beyond those seen in single-agent settings. We present an actor-critic algorithm that trains decentralized policies in multi-agent settings, using centrally computed critics that share an attention mechanism which selects relevant information for each agent at every timestep. This attention mechanism enables more effective and scalable learning in complex multi-agent environments, when compared to recent approaches. Our approach is applicable not only to cooperative settings with shared rewards, but also individualized reward settings, including adversarial settings, as well as settings that do not provide global states, and it makes no assumptions about the action spaces of the agents. As such, it is flexible enough to be applied to most multi-agent learning problems.
****
Complementary-Label Learning for Arbitrary Losses and Models
Takashi Ishida, Gang Niu, Aditya Menon, Masashi Sugiyama
In contrast to the standard classification paradigm where the true class is given to each training pattern, complementary-label learning only uses training patterns each equipped with a complementary label, which only specifies one of the classes that the pattern does not belong to. The goal of this paper is to derive a novel framework of complementary-label learning with an unbiased estimator of the classification risk, for arbitrary losses and models—all existing methods have failed to achieve this goal. Not only is this beneficial for the learning stage, it also makes model/hyper-parameter selection (through cross-validation) possible without the need of any ordinarily labeled validation data, while using any linear/non-linear models or convex/non-convex loss functions. We further improve the risk estimator by a non-negative correction and gradient ascent trick, and demonstrate its superiority through experiments.
****
Causal Identification under Markov Equivalence: Completeness Results
Amin Jaber, Jiji Zhang, Elias Bareinboim
Causal effect identification is the task of determining whether a causal distribution is computable from the combination of an observational distribution and substantive knowledge about the domain under investigation. One of the most studied versions of this problem assumes that knowledge is articulated in the form of a fully known causal diagram, which is arguably a strong assumption in many settings. In this paper, we relax this requirement and consider that the knowledge is articulated in the form of an equivalence class of causal diagrams, in particular, a partial ancestral graph (PAG). This is attractive because a PAG can be learned directly from data, and the scientist does not need to commit to a particular, unique diagram. There are different sufficient conditions for identification in PAGs, but none is complete. We derive a complete algorithm for identification given a PAG. This implies that whenever the causal effect is identifiable, the algorithm returns a valid identification expression; alternatively, it will throw a failure condition, which means that the effect is provably not identifiable. We further provide a graphical characterization of non-identifiability of cau

sal effects in PAGs.
****
Learning from a Learner
Alexis Jacq, Matthieu Geist, Ana Paiva, Olivier Pietquin
In this paper, we propose a novel setting for Inverse Reinforcement Learning (IRL), namely "Learning from a Learner" (LfL). As opposed to standard IRL, it does not consist in learning a reward by observing an optimal agent but from observations of another learning (and thus sub-optimal) agent. To do so, we leverage the fact that the observed agent's policy is assumed to improve over time. The ultimate goal of this approach is to recover the actual environment's reward and to allow the observer to outperform the learner. To recover that reward in practice, we propose methods based on the entropy-regularized policy iteration framework. We discuss different approaches to learn solely from trajectories in the state-action space. We demonstrate the genericity of our method by observing agents implementing various reinforcement learning algorithms. Finally, we show that, on both discrete and continuous state/action tasks, the observer's performance (that optimizes the recovered reward) can surpass those of the observed agent.
****
Differentially Private Fair Learning
Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi -Malvajerdi, Jonathan Ullman
Motivated by settings in which predictive models may be required to be non-discriminatory with respect to certain attributes (such as race), but even collecting the sensitive attribute may be forbidden or restricted, we initiate the study of fair learning under the constraint of differential privacy. Our first algorithm is a private implementation of the equalized odds post-processing approach of (Hardt et al., 2016). This algorithm is appealingly simple, but must be able to use protected group membership explicitly at test time, which can be viewed as a form of “disparate treatment”. Our second algorithm is a differentially private version of the oracle-efficient in-processing approach of (Agarwal et al., 2018) which is more complex but need not have access to protected group membership at test time. We identify new tradeoffs between fairness, accuracy, and privacy that emerge only when requiring all three properties, and show that these tradeoffs can be milder if group membership may be used at test time. We conclude with a brief experimental evaluation.
****
Sum-of-Squares Polynomial Flow
Priyank Jaini, Kira A. Selby, Yaoliang Yu
Triangular map is a recent construct in probability theory that allows one to transform any source probability density function to any target density function. Based on triangular maps, we propose a general framework for high-dimensional density estimation, by specifying one-dimensional transformations (equivalently conditional densities) and appropriate conditioner networks. This framework (a) reveals the commonalities and differences of existing autoregressive and flow based methods, (b) allows a unified understanding of the limitations and representation power of these recent approaches and, (c) motivates us to uncover a new Sum-of-Squares (SOS) flow that is interpretable, universal, and easy to train. We perform several synthetic experiments on various density geometries to demonstrate the benefits (and short-comings) of such transformations. SOS flows achieve competitive results in simulations and several real-world datasets.
****
DBSCAN++: Towards fast and scalable density clustering
Jennifer Jang, Heinrich Jiang
DBSCAN is a classical density-based clustering procedure with tremendous practical relevance. However, DBSCAN implicitly needs to compute the empirical density for each sample point, leading to a quadratic worst-case time complexity, which is too slow on large datasets. We propose DBSCAN++, a simple modification of DBSCAN which only requires computing the densities for a chosen subset of points. We show empirically that, compared to traditional DBSCAN, DBSCAN++ can provide not only competitive performance but also added robustness in the bandwidth hyperp

arameter while taking a fraction of the runtime. We also present statistical consistency guarantees showing the trade-off between computational cost and estimation rates. Surprisingly, up to a certain point, we can enjoy the same estimation rates while lowering computational cost, showing that DBSCAN++ is a sub-quadratic algorithm that attains minimax optimal rates for level-set estimation, a quality that may be of independent interest.
****

Learning What and Where to Transfer

Yunhun Jang, Hankook Lee, Sung Ju Hwang, Jinwoo Shin

As the application of deep learning has expanded to real-world problems with insufficient volume of training data, transfer learning recently has gained much attention as means of improving the performance in such small-data regime. However, when existing methods are applied between heterogeneous architectures and tasks, it becomes more important to manage their detailed configurations and often requires exhaustive tuning on them for the desired performance. To address the issue, we propose a novel transfer learning approach based on meta-learning that can automatically learn what knowledge to transfer from the source network to where in the target network. Given source and target networks, we propose an efficient training scheme to learn meta-networks that decide (a) which pairs of layers between the source and target networks should be matched for knowledge transfer and (b) which features and how much knowledge from each feature should be transferred. We validate our meta-transfer approach against recent transfer learning methods on various datasets and network architectures, on which our automated scheme significantly outperforms the prior baselines that find “what and where to transfer” in a hand-crafted manner.
****

Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, Dj Strouse, Joel Z. Leibo, Nando De Freitas

We propose a unified mechanism for achieving coordination and communication in Multi-Agent Reinforcement Learning (MARL), through rewarding agents for having causal influence over other agents’ actions. Causal influence is assessed using counterfactual reasoning. At each timestep, an agent simulates alternate actions that it could have taken, and computes their effect on the behavior of other agents. Actions that lead to bigger changes in other agents’ behavior are considered influential and are rewarded. We show that this is equivalent to rewarding agents for having high mutual information between their actions. Empirical results demonstrate that influence leads to enhanced coordination and communication in challenging social dilemma environments, dramatically increasing the learning curves of the deep RL agents, and leading to more meaningful learned communication protocols. The influence rewards for all agents can be computed in a decentralized way by enabling agents to learn a model of other agents using deep neural networks. In contrast, key previous works on emergent communication in the MARL setting were unable to learn diverse policies in a decentralized manner and had to resort to centralized training. Consequently, the influence reward opens up a window of new opportunities for research in this area.
****

A Deep Reinforcement Learning Perspective on Internet Congestion Control

Nathan Jay, Noga Rotman, Brighten Godfrey, Michael Schapira, Aviv Tamar

We present and investigate a novel and timely application domain for deep reinforcement learning (RL): Internet congestion control. Congestion control is the core networking task of modulating traffic sources’ data-transmission rates to efficiently utilize network capacity, and is the subject of extensive attention in light of the advent of Internet services such as live video, virtual reality, Internet-of-Things, and more. We show that casting congestion control as RL enables training deep network policies that capture intricate patterns in data traffic and network conditions, and leverage this to outperform the state-of-the-art. We also highlight significant challenges facing real-world adoption of RL-based congestion control, including fairness, safety, and generalization, which are not

trivial to address within conventional RL formalism. To facilitate further research and reproducibility of our results, we present a test suite for RL-guided congestion control based on the OpenAI Gym interface.
****

Graph Neural Network for Music Score Data and Modeling Expressive Piano Performance

Dasaem Jeong, Taegyun Kwon, Yoojin Kim, Juhan Nam

Music score is often handled as one-dimensional sequential data. Unlike words in a text document, notes in music score can be played simultaneously by the polyphonic nature and each of them has its own duration. In this paper, we represent the unique form of musical score using graph neural network and apply it for rendering expressive piano performance from the music score. Specifically, we design the model using note-level gated graph neural network and measure-level hierarchical attention network with bidirectional long short-term memory with an iterative feedback method. In addition, to model different styles of performance for a given input score, we employ a variational auto-encoder. The result of the listening test shows that our proposed model generated more human-like performances compared to a baseline model and a hierarchical attention network model that handles music score as a word-like sequence.
****

Ladder Capsule Network

Taewon Jeong, Youngmin Lee, Heeyoung Kim

We propose a new architecture of the capsule network called the ladder capsule network, which has an alternative building block to the dynamic routing algorithm in the capsule network (Sabour et al., 2017). Motivated by the need for using only important capsules during training for robust performance, we first introduce a new layer called the pruning layer, which removes irrelevant capsules. Based on the selected capsules, we construct higher-level capsule outputs. Subsequently, to capture the part-whole spatial relationships, we introduce another new layer called the ladder layer, the outputs of which are regressed lower-level capsule outputs from higher-level capsules. Unlike the capsule network adopting the routing-by-agreement, the ladder capsule network uses backpropagation from a loss function to reconstruct the lower-level capsule outputs from higher-level capsules; thus, the ladder layer implements the reverse directional inference of the agreement/disagreement mechanism of the capsule network. The experiments on MNIST demonstrate that the ladder capsule network learns an equivariant representation and improves the capability to extrapolate or generalize to pose variations.
****

Training CNNs with Selective Allocation of Channels

Jongheon Jeong, Jinwoo Shin

Recent progress in deep convolutional neural networks (CNNs) have enabled a simple paradigm of architecture design: larger models typically achieve better accuracy. Due to this, in modern CNN architectures, it becomes more important to design models that generalize well under certain resource constraints, e.g. the number of parameters. In this paper, we propose a simple way to improve the capacity of any CNN model having large-scale features, without adding more parameters. In particular, we modify a standard convolutional layer to have a new functionality of channel-selectivity, so that the layer is trained to select important channels to re-distribute their parameters. Our experimental results under various CNN architectures and datasets demonstrate that the proposed new convolutional layer allows new optima that generalize better via efficient resource utilization, compared to the baseline.
****

Learning Discrete and Continuous Factors of Data via Alternating Disentanglement

Yeonwoo Jeong, Hyun Oh Song

We address the problem of unsupervised disentanglement of discrete and continuous explanatory factors of data. We first show a simple procedure for minimizing the total correlation of the continuous latent variables without having to use a discriminator network or perform importance sampling, via cascading the information flow in the beta-VAE framework. Furthermore, we propose a method which avoid

s offloading the entire burden of jointly modeling the continuous and discrete factors to the variational encoder by employing a separate discrete inference procedure. This leads to an interesting alternating minimization problem which switches between finding the most likely discrete configuration given the continuous factors and updating the variational encoder based on the computed discrete factors. Experiments show that the proposed method clearly disentangles discrete factors and significantly outperforms current disentanglement methods based on the disentanglement score and inference network classification score. The source code is available at https://github.com/snumllab/DisentanglementICML19.
****
Improved Zeroth-Order Variance Reduced Algorithms and Analysis for Nonconvex Optimization
Kaiyi Ji, Zhe Wang, Yi Zhou, Yingbin Liang
Two types of zeroth-order stochastic algorithms have recently been designed for nonconvex optimization respectively based on the first-order techniques SVRG and SARAH/SPIDER. This paper addresses several important issues that are still open in these methods. First, all existing SVRG-type zeroth-order algorithms suffer from worse function query complexities than either zeroth-order gradient descent (ZO-GD) or stochastic gradient descent (ZO-SGD). In this paper, we propose a new algorithm ZO-SVRG-Coord-Rand and develop a new analysis for an existing ZO-SVRG-Coord algorithm proposed in Liu et al. 2018b, and show that both ZO-SVRG-Coord-Rand and ZO-SVRG-Coord (under our new analysis) outperform other exiting SVRG-type zeroth-order methods as well as ZO-GD and ZO-SGD. Second, the existing SPIDER-type algorithm SPIDER-SZO (Fang et al., 2018) has superior theoretical performance, but suffers from the generation of a large number of Gaussian random variables as well as a $\sqrt{\epsilon}$-level stepsize in practice. In this paper, we develop a new algorithm ZO-SPIDER-Coord, which is free from Gaussian variable generation and allows a large constant stepsize while maintaining the same convergence rate and query complexity, and we further show that ZO-SPIDER-Coord automatically achieves a linear convergence rate as the iterate enters into a local PL region without restart and algorithmic modification.
****
Neural Logic Reinforcement Learning
Zhengyao Jiang, Shan Luo
Deep reinforcement learning (DRL) has achieved significant breakthroughs in various tasks. However, most DRL algorithms suffer a problem of generalising the learned policy, which makes the policy performance largely affected even by minor modifications of the training environment. Except that, the use of deep neural networks makes the learned policies hard to be interpretable. To address these two challenges, we propose a novel algorithm named Neural Logic Reinforcement Learning (NLRL) to represent the policies in reinforcement learning by first-order logic. NLRL is based on policy gradient methods and differentiable inductive logic programming that have demonstrated significant advantages in terms of interpretability and generalisability in supervised tasks. Extensive experiments conducted on cliff-walking and blocks manipulation tasks demonstrate that NLRL can induce interpretable policies achieving near-optimal performance while showing good generalisability to environments of different initial states and problem sizes.
****
Finding Options that Minimize Planning Time
Yuu Jinnai, David Abel, David Hershkowitz, Michael Littman, George Konidaris
We formalize the problem of selecting the optimal set of options for planning as that of computing the smallest set of options so that planning converges in less than a given maximum of value-iteration passes. We first show that the problem is $\NP$-hard, even if the task is constrained to be deterministic—the first such complexity result for option discovery. We then present the first polynomial-time boundedly suboptimal approximation algorithm for this setting, and empirically evaluate it against both the optimal options and a representative collection of heuristic approaches in simple grid-based domains.
****
Discovering Options for Exploration by Minimizing Cover Time

Yuu Jinnai, Jee Won Park, David Abel, George Konidaris

One of the main challenges in reinforcement learning is solving tasks with sparse reward. We show that the difficulty of discovering a distant rewarding state in an MDP is bounded by the expected cover time of a random walk over the graph induced by the MDP’s transition dynamics. We therefore propose to accelerate exploration by constructing options that minimize cover time. We introduce a new option discovery algorithm that diminishes the expected cover time by connecting the most distant states in the state-space graph with options. We show empirically that the proposed algorithm improves learning in several domains with sparse rewards.

****

Kernel Mean Matching for Content Addressability of GANs

Wittawat Jitkrittum, Patsorn Sangkloy, Muhammad Waleed Gondal, Amit Raj, James Hays, Bernhard Schölkopf

We propose a novel procedure which adds "content-addressability" to any given unconditional implicit model e.g., a generative adversarial network (GAN). The procedure allows users to control the generative process by specifying a set (arbitrary size) of desired examples based on which similar samples are generated from the model. The proposed approach, based on kernel mean matching, is applicable to any generative models which transform latent vectors to samples, and does not require retraining of the model. Experiments on various high-dimensional image generation problems (CelebA-HQ, LSUN bedroom, bridge, tower) show that our approach is able to generate images which are consistent with the input set, while retaining the image quality of the original model. To our knowledge, this is the first work that attempts to construct, at test time, a content-addressable generative model from a trained marginal model.

****

GOODE: A Gaussian Off-The-Shelf Ordinary Differential Equation Solver

David John, Vincent Heuveline, Michael Schober

There are two types of ordinary differential equations (ODEs): initial value problems (IVPs) and boundary value problems (BVPs). While many probabilistic numerical methods for the solution of IVPs have been presented to-date, there exists no efficient probabilistic general-purpose solver for nonlinear BVPs. Our method based on iterated Gaussian process (GP) regression returns a GP posterior over the solution of nonlinear ODEs, which provides a meaningful error estimation via its predictive posterior standard deviation. Our solver is fast (typically of quadratic convergence rate) and the theory of convergence can be transferred from prior non-probabilistic work. Our method performs on par with standard codes for an established benchmark of test problems.

****

Bilinear Bandits with Low-rank Structure

Kwang-Sung Jun, Rebecca Willett, Stephen Wright, Robert Nowak

We introduce the bilinear bandit problem with low-rank structure in which an action takes the form of a pair of arms from two different entity types, and the reward is a bilinear function of the known feature vectors of the arms. The unknown in the problem is a d_1 by d_2 matrix $\mathbf{\Theta}^*$ that defines the reward, and has low rank $r \ll \min\{d_1,d_2\}$. Determination of $\mathbf{\Theta}^*$ with this low-rank structure poses a significant challenge in finding the right exploration-exploitation tradeoff. In this work, we propose a new two-stage algorithm called “Explore-Subspace-Then-Refine” (ESTR). The first stage is an explicit subspace exploration, while the second stage is a linear bandit algorithm called “almost-low-dimensional OFUL” (LowOFUL) that exploits and further refines the estimated subspace via a regularization technique. We show that the regret of ESTR is $\widetilde{\mathcal{O}}((d_1+d_2)^{3/2} \sqrt{r T})$ where $\widetilde{\mathcal{O}}$ hides logarithmic factors and T is the time horizon, which improves upon the regret of $\widetilde{\mathcal{O}}(d_1d_2\sqrt{T})$ attained for a naïve linear bandit reduction. We conjecture that the regret bound of ESTR is unimprovable up to polylogarithmic factors, and our preliminary experiment shows that ESTR outperforms a naïve linear bandit reduction.

****

Statistical Foundations of Virtual Democracy
Anson Kahng, Min Kyung Lee, Ritesh Noothigattu, Ariel Procaccia, Christos-Alexandros Psomas
Virtual democracy is an approach to automating decisions, by learning models of the preferences of individual people, and, at runtime, aggregating the predicted preferences of those people on the dilemma at hand. One of the key questions is which aggregation method – or voting rule – to use; we offer a novel statistical viewpoint that provides guidance. Specifically, we seek voting rules that are robust to prediction errors, in that their output on people’s true preferences is likely to coincide with their output on noisy estimates thereof. We prove that the classic Borda count rule is robust in this sense, whereas any voting rule belonging to the wide family of pairwise-majority consistent rules is not. Our empirical results further support, and more precisely measure, the robustness of Borda count.
****
Molecular Hypergraph Grammar with Its Application to Molecular Optimization
Hiroshi Kajino
Molecular optimization aims to discover novel molecules with desirable properties, and its two fundamental challenges are: (i) it is not trivial to generate valid molecules in a controllable way due to hard chemical constraints such as the valency conditions, and (ii) it is often costly to evaluate a property of a novel molecule, and therefore, the number of property evaluations is limited. These challenges are to some extent alleviated by a combination of a variational autoencoder (VAE) and Bayesian optimization (BO), where VAE converts a molecule into/from its latent continuous vector, and BO optimizes a latent continuous vector (and its corresponding molecule) within a limited number of property evaluations. While the most recent work, for the first time, achieved 100% validity, its architecture is rather complex due to auxiliary neural networks other than VAE, making it difficult to train. This paper presents a molecular hypergraph grammar variational autoencoder (MHG-VAE), which uses a single VAE to achieve 100% validity. Our idea is to develop a graph grammar encoding the hard chemical constraints, called molecular hypergraph grammar (MHG), which guides VAE to always generate valid molecules. We also present an algorithm to construct MHG from a set of molecules.
****
Robust Influence Maximization for Hyperparametric Models
Dimitris Kalimeris, Gal Kaplun, Yaron Singer
In this paper we study the problem of robust influence maximization in the independent cascade model under a hyperparametric assumption. In social networks users influence and are influenced by individuals with similar characteristics and as such they are associated with some features. A recent surging research direction in influence maximization focuses on the case where the edge probabilities on the graph are not arbitrary but are generated as a function of the features of the users and a global hyperparameter. We propose a model where the objective is to maximize the worst-case number of influenced users for any possible value of that hyperparameter. We provide theoretical results showing that proper robust solution in our model is NP-hard and an algorithm that achieves improper robust optimization. We make-use of sampling based techniques and of the renowned multiplicative weight updates algorithm. Additionally we validate our method empirically and prove that it outperforms the state-of-the-art robust influence maximization techniques.
****
Classifying Treatment Responders Under Causal Effect Monotonicity
Nathan Kallus
In the context of individual-level causal inference, we study the problem of predicting whether someone will respond or not to a treatment based on their features and past examples of features, treatment indicator (e.g., drug/no drug), and a binary outcome (e.g., recovery from disease). As a classification task, the problem is made difficult by not knowing the example outcomes under the opposite treatment indicators. We assume the effect is monotonic, as in advertising’s effe

ct on a purchase or bail-setting’s effect on reappearance in court: either it would have happened regardless of treatment, not happened regardless, or happened only depending on exposure to treatment. Predicting whether the latter is latently the case is our focus. While previous work focuses on conditional average treatment effect estimation, formulating the problem as a classification task allows us to develop new tools more suited to this problem. By leveraging monotonicity, we develop new discriminative and generative algorithms for the responder-classification problem. We explore and discuss connections to corrupted data and policy learning. We provide an empirical study with both synthetic and real datasets to compare these specialized algorithms to standard benchmarks.
****

Trainable Decoding of Sets of Sequences for Neural Sequence Models
Ashwin Kalyan, Peter Anderson, Stefan Lee, Dhruv Batra
Many sequence prediction tasks admit multiple correct outputs and so, it is often useful to decode a set of outputs that maximize some task-specific set-level metric. However, retooling standard sequence prediction procedures tailored towards predicting the single best output leads to the decoding of sets containing very similar sequences; failing to capture the variation in the output space. To address this, we propose ∇BS, a trainable decoding procedure that outputs a set of sequences, highly valued according to the metric. Our method tightly integrates the training and decoding phases and further allows for the optimization of the task-specific metric addressing the shortcomings of standard sequence prediction. Further, we discuss the trade-offs of commonly used set-level metrics and motivate a new set-level metric that naturally evaluates the notion of “capturing the variation in the output space”. Finally, we show results on the image captioning task and find that our model outperforms standard techniques and natural ablations.
****

Myopic Posterior Sampling for Adaptive Goal Oriented Design of Experiments
Kirthevasan Kandasamy, Willie Neiswanger, Reed Zhang, Akshay Krishnamurthy, Jeff Schneider, Barnabas Poczos
Bayesian methods for adaptive decision-making, such as Bayesian optimisation, active learning, and active search have seen great success in relevant applications. However, real world data collection tasks are more broad and complex, as we may need to achieve a combination of the above goals and/or application specific goals. In such scenarios, specialised methods have limited applicability. In this work, we design a new myopic strategy for a wide class of adaptive design of experiment (DOE) problems, where we wish to collect data in order to fulfil a given goal. Our approach, Myopic Posterior Sampling (MPS), which is inspired by the classical posterior sampling algorithm for multi-armed bandits, enables us to address a broad suite of DOE tasks where a practitioner may incorporate domain expertise about the system and specify her desired goal via a reward function. Empirically, this general-purpose strategy is competitive with more specialised methods in a wide array of synthetic and real world DOE tasks. More importantly, it enables addressing complex DOE goals where no existing method seems applicable. On the theoretical side, we leverage ideas from adaptive submodularity and reinforcement learning to derive conditions under which MPS achieves sublinear regret against natural benchmark policies.
****

Differentially Private Learning of Geometric Concepts
Haim Kaplan, Yishay Mansour, Yossi Matias, Uri Stemmer
We present differentially private efficient algorithms for learning union of polygons in the plane (which are not necessarily convex). Our algorithms achieve (α,β)-PAC learning and (ϵ,δ)-differential privacy using a sample of size $\tilde{O}\left(\frac{1}{\alpha\epsilon}k\log d\right)$, where the domain is $[d]\times[d]$ and k is the number of edges in the union of polygons.
****

Policy Consolidation for Continual Reinforcement Learning
Christos Kaplanis, Murray Shanahan, Claudia Clopath

We propose a method for tackling catastrophic forgetting in deep reinforcement l
earning that is agnostic to the timescale of changes in the distribution of expe
riences, does not require knowledge of task boundaries and can adapt in continuo
usly changing environments. In our policy consolidation model, the policy networ
k interacts with a cascade of hidden networks that simultaneously remember the a
gent’s policy at a range of timescales and regularise the current policy by its
own history, thereby improving its ability to learn without forgetting. We find
that the model improves continual learning relative to baselines on a number of
continuous control tasks in single-task, alternating two-task, and multi-agent c
ompetitive self-play settings.
****
Error Feedback Fixes SignSGD and other Gradient Compression Schemes
Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, Martin Jaggi
Sign-based algorithms (e.g. signSGD) have been proposed as a biased gradient com
pression technique to alleviate the communication bottleneck in training large n
eural networks across multiple workers. We show simple convex counter-examples w
here signSGD does not converge to the optimum. Further, even when it does conver
ge, signSGD may generalize poorly when compared with SGD. These issues arise bec
ause of the biased nature of the sign compression operator. We then show that us
ing error-feedback, i.e. incorporating the error made by the compression operato
r into the next step, overcomes these issues. We prove that our algorithm (EF-SG
D) with arbitrary compression operator achieves the same rate of convergence as
SGD without any additional assumptions. Thus EF-SGD achieves gradient compressio
n for free. Our experiments thoroughly substantiate the theory.
****
Riemannian adaptive stochastic gradient algorithms on matrix manifolds
Hiroyuki Kasai, Pratik Jawanpuria, Bamdev Mishra
Adaptive stochastic gradient algorithms in the Euclidean space have attracted mu
ch attention lately. Such explorations on Riemannian manifolds, on the other han
d, are relatively new, limited, and challenging. This is because of the intrinsi
c non-linear structure of the underlying manifold and the absence of a canonical
 coordinate system. In machine learning applications, however, most manifolds of
 interest are represented as matrices with notions of row and column subspaces.
In addition, the implicit manifold-related constraints may also lie on such subs
paces. For example, the Grassmann manifold is the set of column subspaces. To th
is end, such a rich structure should not be lost by transforming matrices to jus
t a stack of vectors while developing optimization algorithms on manifolds. We p
ropose novel stochastic gradient algorithms for problems on Riemannian matrix ma
nifolds by adapting the row and column subspaces of gradients. Our algorithms ar
e provably convergent and they achieve the convergence rate of order $O(log(T)/s
qrt(T))$, where T is the number of iterations. Our experiments illustrate that
 the proposed algorithms outperform existing Riemannian adaptive stochastic algo
rithms.
****
Neural Inverse Knitting: From Images to Manufacturing Instructions
Alexandre Kaspar, Tae-Hyun Oh, Liane Makatura, Petr Kellnhofer, Wojciech Matusik
Motivated by the recent potential of mass customization brought by whole-garment
 knitting machines, we introduce the new problem of automatic machine instructio
n generation using a single image of the desired physical product, which we appl
y to machine knitting. We propose to tackle this problem by directly learning to
 synthesize regular machine instructions from real images. We create a cured dat
aset of real samples with their instruction counterpart and propose to use synth
etic images to augment it in a novel way. We theoretically motivate our data mix
ing framework and show empirical results suggesting that making real images look
 more synthetic is beneficial in our problem setup.
****
Processing Megapixel Images with Deep Attention-Sampling Models
Angelos Katharopoulos, Francois Fleuret
Existing deep architectures cannot operate on very large signals such as megapix
el images due to computational and memory constraints. To tackle this limitation

, we propose a fully differentiable end-to-end trainable model that samples and processes only a fraction of the full resolution input image. The locations to process are sampled from an attention distribution computed from a low resolution view of the input. We refer to our method as attention sampling and it can process images of several megapixels with a standard single GPU setup. We show that sampling from the attention distribution results in an unbiased estimator of the full model with minimal variance, and we derive an unbiased estimator of the gradient that we use to train our model end-to-end with a normal SGD procedure. This new method is evaluated on three classification tasks, where we show that it allows to reduce computation and memory footprint by an order of magnitude for the same accuracy as classical architectures. We also show the consistency of the sampling that indeed focuses on informative parts of the input images.
****
Robust Estimation of Tree Structured Gaussian Graphical Models
Ashish Katiyar, Jessica Hoffmann, Constantine Caramanis
Consider jointly Gaussian random variables whose conditional independence structure is specified by a graphical model. If we observe realizations of the variables, we can compute the covariance matrix, and it is well known that the support of the inverse covariance matrix corresponds to the edges of the graphical model. Instead, suppose we only have noisy observations. If the noise at each node is independent, we can compute the sum of the covariance matrix and an unknown diagonal. The inverse of this sum is (in general) dense. We ask: can the original independence structure be recovered? We address this question for tree structured graphical models. We prove that this problem is unidentifiable, but show that this unidentifiability is limited to a small class of candidate trees. We further present additional constraints under which the problem is identifiable. Finally, we provide an O(n^3) algorithm to find this equivalence class of trees.
****
Shallow-Deep Networks: Understanding and Mitigating Network Overthinking
Yigitcan Kaya, Sanghyun Hong, Tudor Dumitras
We characterize a prevalent weakness of deep neural networks (DNNs), ’overthinking’, which occurs when a DNN can reach correct predictions before its final layer. Overthinking is computationally wasteful, and it can also be destructive when, by the final layer, a correct prediction changes into a misclassification. Understanding overthinking requires studying how each prediction evolves during a DNN’s forward pass, which conventionally is opaque. For prediction transparency, we propose the Shallow-Deep Network (SDN), a generic modification to off-the-shelf DNNs that introduces internal classifiers. We apply SDN to four modern architectures, trained on three image classification tasks, to characterize the overthinking problem. We show that SDNs can mitigate the wasteful effect of overthinking with confidence-based early exits, which reduce the average inference cost by more than 50% and preserve the accuracy. We also find that the destructive effect occurs for 50% of misclassifications on natural inputs and that it can be induced, adversarially, with a recent backdooring attack. To mitigate this effect, we propose a new confusion metric to quantify the internal disagreements that will likely to lead to misclassifications.
****
Submodular Streaming in All Its Glory: Tight Approximation, Minimum Memory and Low Adaptive Complexity
Ehsan Kazemi, Marko Mitrovic, Morteza Zadimoghaddam, Silvio Lattanzi, Amin Karbasi
Streaming algorithms are generally judged by the quality of their solution, memory footprint, and computational complexity. In this paper, we study the problem of maximizing a monotone submodular function in the streaming setting with a cardinality constraint k. We first propose SIEVE-STREAMING++, which requires just one pass over the data, keeps only $O(k)$ elements and achieves the tight $\frac{1}{2}$-approximation guarantee. The best previously known streaming algorithms either achieve a suboptimal $\frac{1}{4}$-approximation with $\Theta(k)$ memory or the optimal $\frac{1}{2}$-approximation with $O(k\log k)$ memory. Next, we show that by buffering a small fraction of the stream and applying a careful filt

ering procedure, one can heavily reduce the number of adaptive computational rounds, thus substantially lowering the computational complexity of SIEVE-STREAMING++. We then generalize our results to the more challenging multi-source streaming setting. We show how one can achieve the tight $\frac{1}{2}$-approximation guarantee with $O(k)$ shared memory, while minimizing not only the rounds of computations but also the total number of communicated bits. Finally, we demonstrate the efficiency of our algorithms on real-world data summarization tasks for multi-source streams of tweets and of YouTube videos.
****
Adaptive Scale-Invariant Online Algorithms for Learning Linear Models
Michal Kempka, Wojciech Kotlowski, Manfred K. Warmuth
We consider online learning with linear models, where the algorithm predicts on sequentially revealed instances (feature vectors), and is compared against the best linear function (comparator) in hindsight. Popular algorithms in this framework, such as Online Gradient Descent (OGD), have parameters (learning rates), which ideally should be tuned based on the scales of the features and the optimal comparator, but these quantities only become available at the end of the learning process. In this paper, we resolve the tuning problem by proposing online algorithms making predictions which are invariant under arbitrary rescaling of the features. The algorithms have no parameters to tune, do not require any prior knowledge on the scale of the instances or the comparator, and achieve regret bounds matching (up to a logarithmic factor) that of OGD with optimally tuned separate learning rates per dimension, while retaining comparable runtime performance.
****
CHiVE: Varying Prosody in Speech Synthesis with a Linguistically Driven Dynamic Hierarchical Conditional Variational Network
Tom Kenter, Vincent Wan, Chun-An Chan, Rob Clark, Jakub Vit
The prosodic aspects of speech signals produced by current text-to-speech systems are typically averaged over training material, and as such lack the variety and liveliness found in natural speech. To avoid monotony and averaged prosody contours, it is desirable to have a way of modeling the variation in the prosodic aspects of speech, so audio signals can be synthesized in multiple ways for a given text. We present a new, hierarchically structured conditional variational auto-encoder to generate prosodic features (fundamental frequency, energy and duration) suitable for use with a vocoder or a generative model like WaveNet. At inference time, an embedding representing the prosody of a sentence may be sampled from the variational layer to allow for prosodic variation. To efficiently capture the hierarchical nature of the linguistic input (words, syllables and phones), both the encoder and decoder parts of the auto-encoder are hierarchical, in line with the linguistic structure, with layers being clocked dynamically at the respective rates. We show in our experiments that our dynamic hierarchical network outperforms a non-hierarchical state-of-the-art baseline, and, additionally, that prosody transfer across sentences is possible by employing the prosody embedding of one sentence to generate the speech signal of another.
****
Collaborative Evolutionary Reinforcement Learning
Shauharda Khadka, Somdeb Majumdar, Tarek Nassar, Zach Dwiel, Evren Tumer, Santiago Miret, Yinyin Liu, Kagan Tumer
Deep reinforcement learning algorithms have been successfully applied to a range of challenging control tasks. However, these methods typically struggle with achieving effective exploration and are extremely sensitive to the choice of hyperparameters. One reason is that most approaches use a noisy version of their operating policy to explore - thereby limiting the range of exploration. In this paper, we introduce Collaborative Evolutionary Reinforcement Learning (CERL), a scalable framework that comprises a portfolio of policies that simultaneously explore and exploit diverse regions of the solution space. A collection of learners - typically proven algorithms like TD3 - optimize over varying time-horizons leading to this diverse portfolio. All learners contribute to and use a shared replay buffer to achieve greater sample efficiency. Computational resources are dynamically distributed to favor the best learners as a form of online algorithm sele

ction. Neuroevolution binds this entire process to generate a single emergent learner that exceeds the capabilities of any individual learner. Experiments in a range of continuous control benchmarks demonstrate that the emergent learner significantly outperforms its composite learners while remaining overall more sample-efficient - notably solving the Mujoco Humanoid benchmark where all of its composite learners (TD3) fail entirely in isolation.
****
Geometry Aware Convolutional Filters for Omnidirectional Images Representation
Renata Khasanova, Pascal Frossard
Due to their wide field of view, omnidirectional cameras are frequently used by autonomous vehicles, drones and robots for navigation and other computer vision tasks. The images captured by such cameras, are often analyzed and classified with techniques designed for planar images that unfortunately fail to properly handle the native geometry of such images and therefore results in suboptimal performance. In this paper we aim at improving popular deep convolutional neural networks so that they can properly take into account the specific properties of omnidirectional data. In particular we propose an algorithm that adapts convolutional layers, which often serve as a core building block of a CNN, to the properties of omnidirectional images. Thus, our filters have a shape and size that adapt to the location on the omnidirectional image. We show that our method is not limited to spherical surfaces and is able to incorporate the knowledge about any kind of projective geometry inside the deep learning network. As depicted by our experiments, our method outperforms the existing deep neural network techniques for omnidirectional image classification and compression tasks.
****
EMI: Exploration with Mutual Information
Hyoungseok Kim, Jaekyeom Kim, Yeonwoo Jeong, Sergey Levine, Hyun Oh Song
Reinforcement learning algorithms struggle when the reward signal is very sparse. In these cases, naive random exploration methods essentially rely on a random walk to stumble onto a rewarding state. Recent works utilize intrinsic motivation to guide the exploration via generative models, predictive forward models, or discriminative modeling of novelty. We propose EMI, which is an exploration method that constructs embedding representation of states and actions that does not rely on generative decoding of the full observation but extracts predictive signals that can be used to guide exploration based on forward prediction in the representation space. Our experiments show competitive results on challenging locomotion tasks with continuous control and on image-based exploration tasks with discrete actions on Atari. The source code is available at https://github.com/snu-mllab/EMI.
****
FloWaveNet : A Generative Flow for Raw Audio
Sungwon Kim, Sang-Gil Lee, Jongyoon Song, Jaehyeon Kim, Sungroh Yoon
Most modern text-to-speech architectures use a WaveNet vocoder for synthesizing high-fidelity waveform audio, but there have been limitations, such as high inference time, in practical applications due to its ancestral sampling scheme. The recently suggested Parallel WaveNet and ClariNet has achieved real-time audio synthesis capability by incorporating inverse autoregressive flow (IAF) for parallel sampling. However, these approaches require a two-stage training pipeline with a well-trained teacher network and can only produce natural sound by using probability distillation along with heavily-engineered auxiliary loss terms. We propose FloWaveNet, a flow-based generative model for raw audio synthesis. FloWaveNet requires only a single-stage training procedure and a single maximum likelihood loss, without any additional auxiliary terms, and it is inherently parallel due to the characteristics of generative flow. The model can efficiently sample raw audio in real-time, with clarity comparable to previous two-stage parallel models. The code and samples for all models, including our FloWaveNet, are available on GitHub.
****
Curiosity-Bottleneck: Exploration By Distilling Task-Specific Novelty
Youngjin Kim, Wontae Nam, Hyunwoo Kim, Ji-Hoon Kim, Gunhee Kim

Exploration based on state novelty has brought great success in challenging reinforcement learning problems with sparse rewards. However, existing novelty-based strategies become inefficient in real-world problems where observation contains not only task-dependent state novelty of our interest but also task-irrelevant information that should be ignored. We introduce an information- theoretic exploration strategy named Curiosity-Bottleneck that distills task-relevant information from observation. Based on the information bottleneck principle, our exploration bonus is quantified as the compressiveness of observation with respect to the learned representation of a compressive value network. With extensive experiments on static image classification, grid-world and three hard-exploration Atari games, we show that Curiosity-Bottleneck learns an effective exploration strategy by robustly measuring the state novelty in distractive environments where state-of-the-art exploration methods often degenerate.
****
Contextual Multi-armed Bandit Algorithm for Semiparametric Reward Model
Gi-Soo Kim, Myunghee Cho Paik
Contextual multi-armed bandit (MAB) algorithms have been shown promising for maximizing cumulative rewards in sequential decision tasks such as news article recommendation systems, web page ad placement algorithms, and mobile health. However, most of the proposed contextual MAB algorithms assume linear relationships between the reward and the context of the action. This paper proposes a new contextual MAB algorithm for a relaxed, semiparametric reward model that supports nonstationarity. The proposed method is less restrictive, easier to implement and faster than two alternative algorithms that consider the same model, while achieving a tight regret upper bound. We prove that the high-probability upper bound of the regret incurred by the proposed algorithm has the same order as the Thompson sampling algorithm for linear reward models. The proposed and existing algorithms are evaluated via simulation and also applied to Yahoo! news article recommendation log data.
****
Uniform Convergence Rate of the Kernel Density Estimator Adaptive to Intrinsic Volume Dimension
Jisu Kim, Jaehyeok Shin, Alessandro Rinaldo, Larry Wasserman
We derive concentration inequalities for the supremum norm of the difference between a kernel density estimator (KDE) and its point-wise expectation that hold uniformly over the selection of the bandwidth and under weaker conditions on the kernel and the data generating distribution than previously used in the literature. We first propose a novel concept, called the volume dimension, to measure the intrinsic dimension of the support of a probability distribution based on the rates of decay of the probability of vanishing Euclidean balls. Our bounds depend on the volume dimension and generalize the existing bounds derived in the literature. In particular, when the data-generating distribution has a bounded Lebesgue density or is supported on a sufficiently well-behaved lower-dimensional manifold, our bound recovers the same convergence rate depending on the intrinsic dimension of the support as ones known in the literature. At the same time, our results apply to more general cases, such as the ones of distribution with unbounded densities or supported on a mixture of manifolds with different dimensions. Analogous bounds are derived for the derivative of the KDE, of any order. Our results are generally applicable but are especially useful for problems in geometric inference and topological data analysis, including level set estimation, density-based clustering, modal clustering and mode hunting, ridge estimation and persistent homology.
****
Bit-Swap: Recursive Bits-Back Coding for Lossless Compression with Hierarchical Latent Variables
Friso Kingma, Pieter Abbeel, Jonathan Ho
The bits-back argument suggests that latent variable models can be turned into lossless compression schemes. Translating the bits-back argument into efficient and practical lossless compression schemes for general latent variable models, however, is still an open problem. Bits-Back with Asymmetric Numeral Systems (BB-A

NS), recently proposed by Townsend et al,. 2019, makes bits-back coding practica
lly feasible for latent variable models with one latent layer, but it is ineffic
ient for hierarchical latent variable models. In this paper we propose Bit-Swap,
 a new compression scheme that generalizes BB-ANS and achieves strictly better c
ompression rates for hierarchical latent variable models with Markov chain struc
ture. Through experiments we verify that Bit-Swap results in lossless compressio
n rates that are empirically superior to existing techniques.
****
CompILE: Compositional Imitation Learning and Execution
Thomas Kipf, Yujia Li, Hanjun Dai, Vinicius Zambaldi, Alvaro Sanchez-Gonzalez, E
dward Grefenstette, Pushmeet Kohli, Peter Battaglia
We introduce Compositional Imitation Learning and Execution (CompILE): a framewo
rk for learning reusable, variable-length segments of hierarchically-structured
behavior from demonstration data. CompILE uses a novel unsupervised, fully-diffe
rentiable sequence segmentation module to learn latent encodings of sequential d
ata that can be re-composed and executed to perform new tasks. Once trained, our
 model generalizes to sequences of longer length and from environment instances
not seen during training. We evaluate CompILE in a challenging 2D multi-task env
ironment and a continuous control task, and show that it can find correct task b
oundaries and event encodings in an unsupervised manner. Latent codes and associ
ated behavior policies discovered by CompILE can be used by a hierarchical agent
, where the high-level policy selects actions in the latent code space, and the
low-level, task-specific policies are simply the learned decoders. We found that
 our CompILE-based agent could learn given only sparse rewards, where agents wit
hout task-specific policies struggle.
****
Adaptive and Safe Bayesian Optimization in High Dimensions via One-Dimensional S
ubspaces
Johannes Kirschner, Mojmir Mutny, Nicole Hiller, Rasmus Ischebeck, Andreas Kraus
e
Bayesian optimization is known to be difficult to scale to high dimensions, beca
use the acquisition step requires solving a non-convex optimization problem in t
he same search space. In order to scale the method and keep its benefits, we pro
pose an algorithm (LineBO) that restricts the problem to a sequence of iterative
ly chosen one-dimensional sub-problems that can be solved efficiently. We show t
hat our algorithm converges globally and obtains a fast local rate when the func
tion is strongly convex. Further, if the objective has an invariant subspace, ou
r method automatically adapts to the effective dimension without changing the al
gorithm. When combined with the SafeOpt algorithm to solve the sub-problems, we
obtain the first safe Bayesian optimization algorithm with theoretical guarantee
s applicable in high-dimensional settings. We evaluate our method on multiple sy
nthetic benchmarks, where we obtain competitive performance. Further, we deploy
our algorithm to optimize the beam intensity of the Swiss Free Electron Laser wi
th up to 40 parameters while satisfying safe operation constraints.
****
AUCμ: A Performance Metric for Multi-Class Machine Learning Models
Ross Kleiman, David Page
The area under the receiver operating characteristic curve (AUC) is arguably the
 most common metric in machine learning for assessing the quality of a two-class
 classification model. As the number and complexity of machine learning applicat
ions grows, so too does the need for measures that can gracefully extend to clas
sification models trained for more than two classes. Prior work in this area has
 proven computationally intractable and/or inconsistent with known properties of
 AUC, and thus there is still a need for an improved multi-class efficacy metric
. We provide in this work a multi-class extension of AUC that we call AUC{\textm
u} that is derived from first principles of the binary class AUC. AUC{\textmu} h
as similar computational complexity to AUC and maintains the properties of AUC c
ritical to its interpretation and use.
****
Fair k-Center Clustering for Data Summarization

Matthäus Kleindessner, Pranjal Awasthi, Jamie Morgenstern
In data summarization we want to choose k prototypes in order to summarize a d
ata set. We study a setting where the data set comprises several demographic gro
ups and we are restricted to choose k_i prototypes belonging to group i. A c
ommon approach to the problem without the fairness constraint is to optimize a c
entroid-based clustering objective such as k-center. A natural extension then
is to incorporate the fairness constraint into the clustering problem. Existing
algorithms for doing so run in time super-quadratic in the size of the data set,
 which is in contrast to the standard k-center problem being approximable in l
inear time. In this paper, we resolve this gap by providing a simple approximati
on algorithm for the k-center problem under the fairness constraint with runni
ng time linear in the size of the data set and k. If the number of demographic
 groups is small, the approximation guarantee of our algorithm only incurs a con
stant-factor overhead.
****
Guarantees for Spectral Clustering with Fairness Constraints
Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, Jamie Morgenstern
Given the widespread popularity of spectral clustering (SC) for partitioning gra
ph data, we study a version of constrained SC in which we try to incorporate the
 fairness notion proposed by Chierichetti et al. (2017). According to this notio
n, a clustering is fair if every demographic group is approximately proportional
ly represented in each cluster. To this end, we develop variants of both normali
zed and unnormalized constrained SC and show that they help find fairer clusteri
ngs on both synthetic and real data. We also provide a rigorous theoretical anal
ysis of our algorithms on a natural variant of the stochastic block model, where
 h groups have strong inter-group connectivity, but also exhibit a “natural” c
lustering structure which is fair. We prove that our algorithms can recover this
 fair clustering with high probability.
****
POPQORN: Quantifying Robustness of Recurrent Neural Networks
Ching-Yun Ko, Zhaoyang Lyu, Lily Weng, Luca Daniel, Ngai Wong, Dahua Lin
The vulnerability to adversarial attacks has been a critical issue for deep neur
al networks. Addressing this issue requires a reliable way to evaluate the robus
tness of a network. Recently, several methods have been developed to compute rob
ustness quantification for neural networks, namely, certified lower bounds of th
e minimum adversarial perturbation. Such methods, however, were devised for feed
-forward networks, e.g. multi-layer perceptron or convolutional networks. It rem
ains an open problem to quantify robustness for recurrent networks, especially L
STM and GRU. For such networks, there exist additional challenges in computing t
he robustness quantification, such as handling the inputs at multiple steps and
the interaction between gates and states. In this work, we propose POPQORN (Prop
agated-output Quantified Robustness for RNNs), a general algorithm to quantify r
obustness of RNNs, including vanilla RNNs, LSTMs, and GRUs. We demonstrate its e
ffectiveness on different network architectures and show that the robustness qua
ntification on individual steps can lead to new insights.
****
Decentralized Stochastic Optimization and Gossip Algorithms with Compressed Comm
unication
Anastasia Koloskova, Sebastian Stich, Martin Jaggi
We consider decentralized stochastic optimization with the objective function (e
.g. data samples for machine learning tasks) being distributed over n machines t
hat can only communicate to their neighbors on a fixed communication graph. To a
ddress the communication bottleneck, the nodes compress (e.g. quantize or sparsi
fy) their model updates. We cover both unbiased and biased compression operators
 with quality denoted by \delta <= 1 (\delta=1 meaning no compression). We (i) p
ropose a novel gossip-based stochastic gradient descent algorithm, CHOCO-SGD, th
at converges at rate O(1/(nT) + 1/(T \rho^2 \delta)^2) for strongly convex objec
tives, where T denotes the number of iterations and \rho the eigengap of the con
nectivity matrix. We (ii) present a novel gossip algorithm, CHOCO-GOSSIP, for th
e average consensus problem that converges in time O(1/(\rho^2\delta) \log (1/\e

psilon)) for accuracy \epsilon > 0. This is (up to our knowledge) the first gossip algorithm that supports arbitrary compressed messages for \delta > 0 and still exhibits linear convergence. We (iii) show in experiments that both of our algorithms do outperform the respective state-of-the-art baselines and CHOCO-SGD can reduce communication by at least two orders of magnitudes.
****
Robust Learning from Untrusted Sources
Nikola Konstantinov, Christoph Lampert
Modern machine learning methods often require more data for training than a single expert can provide. Therefore, it has become a standard procedure to collect data from multiple external sources, \eg via crowdsourcing. Unfortunately, the quality of these sources is not always guaranteed. As further complications, the data might be stored in a distributed way, or might even have to remain private. In this work, we address the question of how to learn robustly in such scenarios. Studying the problem through the lens of statistical learning theory, we derive a procedure that allows for learning from all available sources, yet automatically suppresses irrelevant or corrupted data. We show by extensive experiments that our method provides significant improvements over alternative approaches from robust statistics and distributed optimization.
****
Stochastic Beams and Where To Find Them: The Gumbel-Top-k Trick for Sampling Sequences Without Replacement
Wouter Kool, Herke Van Hoof, Max Welling
The well-known Gumbel-Max trick for sampling from a categorical distribution can be extended to sample k elements without replacement. We show how to implicitly apply this 'Gumbel-Top-k' trick on a factorized distribution over sequences, allowing to draw exact samples without replacement using a Stochastic Beam Search. Even for exponentially large domains, the number of model evaluations grows only linear in k and the maximum sampled sequence length. The algorithm creates a theoretical connection between sampling and (deterministic) beam search and can be used as a principled intermediate alternative. In a translation task, the proposed method compares favourably against alternatives to obtain diverse yet good quality translations. We show that sequences sampled without replacement can be used to construct low-variance estimators for expected sentence-level BLEU score and model entropy.
****
LIT: Learned Intermediate Representation Training for Model Compression
Animesh Koratana, Daniel Kang, Peter Bailis, Matei Zaharia
Researchers have proposed a range of model compression techniques to reduce the computational and memory footprint of deep neural networks (DNNs). In this work, we introduce Learned Intermediate representation Training (LIT), a novel model compression technique that outperforms a range of recent model compression techniques by leveraging the highly repetitive structure of modern DNNs (e.g., ResNet). LIT uses a teacher DNN to train a student DNN of reduced depth by leveraging two key ideas: 1) LIT directly compares intermediate representations of the teacher and student model and 2) LIT uses the intermediate representation from the teacher model's previous block as input to the current student block during training, improving stability of intermediate representations in the student network. We show that LIT can substantially reduce network size without loss in accuracy on a range of DNN architectures and datasets. For example, LIT can compress ResNet on CIFAR10 by 3.4$\times$ outperforming network slimming and FitNets. Furthermore, LIT can compress, by depth, ResNeXt 5.5$\times$ on CIFAR10 (image classification), VDCNN by 1.7$\times$ on Amazon Reviews (sentiment analysis), and StarGAN by 1.8$\times$ on CelebA (style transfer, i.e., GANs).
****
Similarity of Neural Network Representations Revisited
Simon Kornblith, Mohammad Norouzi, Honglak Lee, Geoffrey Hinton
Recent work has sought to understand the behavior of neural networks by comparing representations between layers and between different trained models. We examine methods for comparing neural network representations based on canonical correl

ation analysis (CCA). We show that CCA belongs to a family of statistics for mea
suring multivariate similarity, but that neither CCA nor any other statistic tha
t is invariant to invertible linear transformation can measure meaningful simila
rities between representations of higher dimension than the number of data point
s. We introduce a similarity index that measures the relationship between repres
entational similarity matrices and does not suffer from this limitation. This si
milarity index is equivalent to centered kernel alignment (CKA) and is also clos
ely connected to CCA. Unlike CCA, CKA can reliably identify correspondences betw
een representations in networks trained from different initializations.
****
On the Complexity of Approximating Wasserstein Barycenters
Alexey Kroshnin, Nazarii Tupitsa, Darina Dvinskikh, Pavel Dvurechensky, Alexande
r Gasnikov, Cesar Uribe
We study the complexity of approximating the Wasserstein barycenter of m discr
ete measures, or histograms of size n, by contrasting two alternative approach
es that use entropic regularization. The first approach is based on the Iterativ
e Bregman Projections (IBP) algorithm for which our novel analysis gives a compl
exity bound proportional to ${mn^2}/{\varepsilon^2}$ to approximate the original
 non-regularized barycenter. On the other hand, using an approach based on accel
erated gradient descent, we obtain a complexity proportional to ${mn^{2}}/{\vare
psilon}$. As a byproduct, we show that the regularization parameter in both appr
oaches has to be proportional to ε, which causes instability of both
 algorithms when the desired accuracy is high. To overcome this issue, we propos
e a novel proximal-IBP algorithm, which can be seen as a proximal gradient metho
d, which uses IBP on each iteration to make a proximal step. We also consider th
e question of scalability of these algorithms using approaches from distributed
optimization and show that the first algorithm can be implemented in a centraliz
ed distributed setting (master/slave), while the second one is amenable to a mor
e general decentralized distributed setting with an arbitrary network topology.
****
Estimate Sequences for Variance-Reduced Stochastic Composite Optimization
Andrei Kulunchakov, Julien Mairal
In this paper, we propose a unified view of gradient-based algorithms for stocha
stic convex composite optimization by extending the concept of estimate sequence
 introduced by Nesterov. This point of view covers the stochastic gradient desce
nt method, variants of the approaches SAGA, SVRG, and has several advantages: (i
) we provide a generic proof of convergence for the aforementioned methods; (ii)
 we show that this SVRG variant is adaptive to strong convexity; (iii) we natura
lly obtain new algorithms with the same guarantees; (iv) we derive generic strat
egies to make these algorithms robust to stochastic noise, which is useful when
data is corrupted by small random perturbations. Finally, we show that this view
point is useful to obtain new accelerated algorithms in the sense of Nesterov.
****
Faster Algorithms for Binary Matrix Factorization
Ravi Kumar, Rina Panigrahy, Ali Rahimi, David Woodruff
We give faster approximation algorithms for well-studied variants of Binary Matr
ix Factorization (BMF), where we are given a binary $m \times n$ matrix A and
would like to find binary rank-k matrices U, V to minimize the Frobenius nor
m of $U \cdot V - A$. In the first setting, $U \cdot V$ denotes multiplication o
ver $\mathbb{Z}$, and we give a constant-factor approximation algorithm that run
s in $2^{O(k^2 \log k)} \textrm{poly}(mn)$ time, improving upon the previous $\m
in(2^{2^k}, 2^n) \textrm{poly}(mn)$ time. Our techniques generalize to minimizin
g $\|U \cdot V - A\|_p$ for $p \geq 1$, in $2^{O(k^{\lceil p/2 \rceil + 1}\log k
)} \textrm{poly}(mn)$ time. For $p = 1$, this has a graph-theoretic consequence,
 namely, a $2^{O(k^2)} \poly(mn)$-time algorithm to approximate a graph as a uni
on of disjoint bicliques. In the second setting, $U \cdot V$ is over $\GF(2)$, a
nd we give a bicriteria constant-factor approximation algorithm that runs in $2^
{O(k^3)} \poly(mn)$ time to find binary rank-$O(k \log m)$ matrices U, V who
se cost is as good as the best rank-k approximation, improving upon $\min(2^{2
^k}mn, \min(m,n)^{k^{O(1)}} \textrm{poly}(mn))$ time.

****
Loss Landscapes of Regularized Linear Autoencoders
Daniel Kunin, Jonathan Bloom, Aleksandrina Goeva, Cotton Seed
Autoencoders are a deep learning model for representation learning. When trained
 to minimize the distance between the data and its reconstruction, linear autoen
coders (LAEs) learn the subspace spanned by the top principal directions but can
not learn the principal directions themselves. In this paper, we prove that $L_2
$-regularized LAEs are symmetric at all critical points and learn the principal
directions as the left singular vectors of the decoder. We smoothly parameterize
 the critical manifold and relate the minima to the MAP estimate of probabilisti
c PCA. We illustrate these results empirically and consider implications for PCA
 algorithms, computational neuroscience, and the algebraic topology of learning.
****
Geometry and Symmetry in Short-and-Sparse Deconvolution
Han-Wen Kuo, Yenson Lau, Yuqian Zhang, John Wright
We study the Short-and-Sparse (SaS) deconvolution problem of recovering a short
signal a0 and a sparse signal x0 from their convolution. We propose a method bas
ed on nonconvex optimization, which under certain conditions recovers the target
 short and sparse signals, up to a signed shift symmetry which is intrinsic to t
his model. This symmetry plays a central role in shaping the optimization landsc
ape for deconvolution. We give a regional analysis, which characterizes this lan
dscape geometrically, on a union of subspaces. Our geometric characterization ho
lds when the length-p0 short signal a0 has shift coherence {\textmu}, and x0 fol
lows a random sparsity model with sparsity rate θ $\in$ [c1/p0, c2/(p0\sq
rt{\mu}+\sqrt{p0})] / (log^2(p0)) . Based on this geometry, we give a provable m
ethod that successfully solves SaS deconvolution with high probability.
****
A Large-Scale Study on Regularization and Normalization in GANs
Karol Kurach, Mario Lu■i■, Xiaohua Zhai, Marcin Michalski, Sylvain Gelly
Generative adversarial networks (GANs) are a class of deep generative models whi
ch aim to learn a target distribution in an unsupervised fashion. While they wer
e successfully applied to many problems, training a GAN is a notoriously challen
ging task and requires a significant number of hyperparameter tuning, neural arc
hitecture engineering, and a non-trivial amount of “tricks". The success in many
 practical applications coupled with the lack of a measure to quantify the failu
re modes of GANs resulted in a plethora of proposed losses, regularization and n
ormalization schemes, as well as neural architectures. In this work we take a so
ber view of the current state of GANs from a practical perspective. We discuss a
nd evaluate common pitfalls and reproducibility issues, open-source our code on
Github, and provide pre-trained models on TensorFlow Hub.
****
Making Decisions that Reduce Discriminatory Impacts
Matt Kusner, Chris Russell, Joshua Loftus, Ricardo Silva
As machine learning algorithms move into real-world settings, it is crucial to e
nsure they are aligned with societal values. There has been much work on one asp
ect of this, namely the discriminatory prediction problem: How can we reduce dis
crimination in the predictions themselves? While an important question, solution
s to this problem only apply in a restricted setting, as we have full control ov
er the predictions. Often we care about the non-discrimination of quantities we
do not have full control over. Thus, we describe another key aspect of this chal
lenge, the discriminatory impact problem: How can we reduce discrimination arisi
ng from the real-world impact of decisions? To address this, we describe causal
methods that model the relevant parts of the real-world system in which the deci
sions are made. Unlike previous approaches, these models not only allow us to ma
p the causal pathway of a single decision, but also to model the effect of inter
ference–how the impact on an individual depends on decisions made about other pe
ople. Often, the goal of decision policies is to maximize a beneficial impact ov
erall. To reduce the discrimination of these benefits, we devise a constraint in
spired by recent work in counterfactual fairness, and give an efficient procedur
e to solve the constrained optimization problem. We demonstrate our approach wit

h an example: how to increase students taking college entrance exams in New York City public schools.
****
Garbage In, Reward Out: Bootstrapping Exploration in Multi-Armed Bandits
Branislav Kveton, Csaba Szepesvari, Sharan Vaswani, Zheng Wen, Tor Lattimore, Mohammad Ghavamzadeh
We propose a bandit algorithm that explores by randomizing its history of rewards. Specifically, it pulls the arm with the highest mean reward in a non-parametric bootstrap sample of its history with pseudo rewards. We design the pseudo rewards such that the bootstrap mean is optimistic with a sufficiently high probability. We call our algorithm Giro, which stands for garbage in, reward out. We analyze Giro in a Bernoulli bandit and derive a $O(K \Delta^{-1} \log n)$ bound on its n-round regret, where Δ is the difference in the expected rewards of the optimal and the best suboptimal arms, and K is the number of arms. The main advantage of our exploration design is that it easily generalizes to structured problems. To show this, we propose contextual Giro with an arbitrary reward generalization model. We evaluate Giro and its contextual variant on multiple synthetic and real-world problems, and observe that it performs well.
****
Characterizing Well-Behaved vs. Pathological Deep Neural Networks
Antoine Labatie
We introduce a novel approach, requiring only mild assumptions, for the characterization of deep neural networks at initialization. Our approach applies both to fully-connected and convolutional networks and easily incorporates batch normalization and skip-connections. Our key insight is to consider the evolution with depth of statistical moments of signal and noise, thereby characterizing the presence or absence of pathologies in the hypothesis space encoded by the choice of hyperparameters. We establish: (i) for feedforward networks, with and without batch normalization, the multiplicativity of layer composition inevitably leads to ill-behaved moments and pathologies; (ii) for residual networks with batch normalization, on the other hand, skip-connections induce power-law rather than exponential behaviour, leading to well-behaved moments and no pathology.
****
State-Reification Networks: Improving Generalization by Modeling the Distribution of Hidden Representations
Alex Lamb, Jonathan Binas, Anirudh Goyal, Sandeep Subramanian, Ioannis Mitliagkas, Yoshua Bengio, Michael Mozer
Machine learning promises methods that generalize well from finite labeled data. However, the brittleness of existing neural net approaches is revealed by notable failures, such as the existence of adversarial examples that are misclassified despite being nearly identical to a training example, or the inability of recurrent sequence-processing nets to stay on track without teacher forcing. We introduce a method, which we refer to as _state reification_, that involves modeling the distribution of hidden states over the training data and then projecting hidden states observed during testing toward this distribution. Our intuition is that if the network can remain in a familiar manifold of hidden space, subsequent layers of the net should be well trained to respond appropriately. We show that this state-reification method helps neural nets to generalize better, especially when labeled data are sparse, and also helps overcome the challenge of achieving robust generalization with adversarial training.
****
A Recurrent Neural Cascade-based Model for Continuous-Time Diffusion
Sylvain Lamprier
Many works have been proposed in the literature to capture the dynamics of diffusion in networks. While some of them define graphical Markovian models to extract temporal relationships between node infections in networks, others consider diffusion episodes as sequences of infections via recurrent neural models. In this paper we propose a model at the crossroads of these two extremes, which embeds the history of diffusion in infected nodes as hidden continuous states. Depending on the trajectory followed by the content before reaching a given node, the di

stribution of influence probabilities may vary. However, content trajectories are usually hidden in the data, which induces challenging learning problems. We propose a topological recurrent neural model which exhibits good experimental performances for diffusion modeling and prediction.
****
Projection onto Minkowski Sums with Application to Constrained Learning
Joong-Ho Won, Jason Xu, Kenneth Lange
We introduce block descent algorithms for projecting onto Minkowski sums of sets. Projection onto such sets is a crucial step in many statistical learning problems, and may regularize complexity of solutions to an optimization problem or arise in dual formulations of penalty methods. We show that projecting onto the Minkowski sum admits simple, efficient algorithms when complications such as overlapping constraints pose challenges to existing methods. We prove that our algorithm converges linearly when sets are strongly convex or satisfy an error bound condition, and extend the theory and methods to encompass non-convex sets as well. We demonstrate empirical advantages in runtime and accuracy over competitors in applications to $\ell_{1,p}$-regularized learning, constrained lasso, and overlapping group lasso.
****
Safe Policy Improvement with Baseline Bootstrapping
Romain Laroche, Paul Trichelair, Remi Tachet Des Combes
This paper considers Safe Policy Improvement (SPI) in Batch Reinforcement Learning (Batch RL): from a fixed dataset and without direct access to the true environment, train a policy that is guaranteed to perform at least as well as the baseline policy used to collect the data. ■ Our approach, called SPI with Baseline Bootstrapping (SPIBB), is inspired by the knows-what-it-knows paradigm: it bootstraps the trained policy with the baseline when the uncertainty is high. ■ Our first algorithm, Π_b-SPIBB, comes with SPI theoretical guarantees. ■ We also implement a variant, $\Pi_{\leq b}$-SPIBB, that is even more efficient in practice. ■ We apply our algorithms to a motivational stochastic gridworld domain and further demonstrate on randomly generated MDPs the superiority of SPIBB with respect to existing algorithms, not only in safety but also in mean performance. ■ Finally, we implement a model-free version of SPIBB and show its benefits on a navigation task with deep RL implementation called SPIBB-DQN, which is, to the best of our knowledge, the first RL algorithm relying on a neural network representation able to train efficiently and reliably from batch data, without any interaction with the environment.
****
A Better k-means++ Algorithm via Local Search
Silvio Lattanzi, Christian Sohler
In this paper, we develop a new variant of k-means++ seeding that in expectation achieves a constant approximation guarantee. We obtain this result by a simple combination of k-means++ sampling with a local search strategy. We evaluate our algorithm empirically and show that it also improves the quality of a solution in practice.
****
Lorentzian Distance Learning for Hyperbolic Representations
Marc Law, Renjie Liao, Jake Snell, Richard Zemel
We introduce an approach to learn representations based on the Lorentzian distance in hyperbolic geometry. Hyperbolic geometry is especially suited to hierarchically-structured datasets, which are prevalent in the real world. Current hyperbolic representation learning methods compare examples with the Poincaré distance. They try to minimize the distance of each node in a hierarchy with its descendants while maximizing its distance with other nodes. This formulation produces node representations close to the centroid of their descendants. To obtain efficient and interpretable algorithms, we exploit the fact that the centroid w.r.t the squared Lorentzian distance can be written in closed-form. We show that the Euclidean norm of such a centroid decreases as the curvature of the hyperbolic space decreases. This property makes it appropriate to represent hierarchies where parent nodes minimize the distances to their descendants and have smaller Euclid

ean norm than their children. Our approach obtains state-of-the-art results in retrieval and classification tasks on different datasets.
****
DP-GP-LVM: A Bayesian Non-Parametric Model for Learning Multivariate Dependency Structures
Andrew Lawrence, Carl Henrik Ek, Neill Campbell
We present a non-parametric Bayesian latent variable model capable of learning dependency structures across dimensions in a multivariate setting. Our approach is based on flexible Gaussian process priors for the generative mappings and interchangeable Dirichlet process priors to learn the structure. The introduction of the Dirichlet process as a specific structural prior allows our model to circumvent issues associated with previous Gaussian process latent variable models. Inference is performed by deriving an efficient variational bound on the marginal log-likelihood of the model. We demonstrate the efficacy of our approach via analysis of discovered structure and superior quantitative performance on missing data imputation.
****
POLITEX: Regret Bounds for Policy Iteration using Expert Prediction
Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, Gellert Weisz
We present POLITEX (POLicy ITeration with EXpert advice), a variant of policy iteration where each policy is a Boltzmann distribution over the sum of action-value function estimates of the previous policies, and analyze its regret in continuing RL problems. We assume that the value function error after running a policy for τ time steps scales as $\epsilon(\tau) = \epsilon_0 + O(\sqrt{d/\tau})$, where ϵ_0 is the worst-case approximation error and d is the number of features in a compressed representation of the state-action space. We establish that this condition is satisfied by the LSPE algorithm under certain assumptions on the MDP and policies. Under the error assumption, we show that the regret of POLITEX in uniformly mixing MDPs scales as $O(d^{1/2}T^{3/4} + \epsilon_0T)$, where $O(\cdot)$ hides logarithmic terms and problem-dependent constants. Thus, we provide the first regret bound for a fully practical model-free method which only scales in the number of features, and not in the size of the underlying MDP. Experiments on a queuing problem confirm that POLITEX is competitive with some of its alternatives, while preliminary results on Ms Pacman (one of the standard Atari benchmark problems) confirm the viability of POLITEX beyond linear function approximation.
****
Batch Policy Learning under Constraints
Hoang Le, Cameron Voloshin, Yisong Yue
When learning policies for real-world domains, two important questions arise: (i) how to efficiently use pre-collected off-policy, non-optimal behavior data; and (ii) how to mediate among different competing objectives and constraints. We thus study the problem of batch policy learning under multiple constraints, and offer a systematic solution. We first propose a flexible meta-algorithm that admits any batch reinforcement learning and online learning procedure as subroutines. We then present a specific algorithmic instantiation and provide performance guarantees for the main objective and all constraints. As part of off-policy learning, we propose a simple method for off-policy policy evaluation (OPE) and derive PAC-style bounds. Our algorithm achieves strong empirical results in different domains, including in a challenging problem of simulated car driving subject to multiple constraints such as lane keeping and smooth driving. We also show experimentally that our OPE method outperforms other popular OPE techniques on a standalone basis, especially in a high-dimensional setting.
****
Target-Based Temporal-Difference Learning
Donghwan Lee, Niao He
The use of target networks has been a popular and key component of recent deep Q-learning algorithms for reinforcement learning, yet little is known from the theory side. In this work, we introduce a new family of target-based temporal diff

erence (TD) learning algorithms that maintain two separate learning parameters {
–} the target variable and online variable. We propose three members in the fami
ly, the averaging TD, double TD, and periodic TD, where the target variable is u
pdated through an averaging, symmetric, or periodic fashion, respectively, mirro
ring those techniques used in deep Q-learning practice. We establish asymptotic
convergence analyses for both averaging TD and double TD and a finite sample ana
lysis for periodic TD. In addition, we provide some simulation results showing p
otentially superior convergence of these target-based TD algorithms compared to
the standard TD-learning. While this work focuses on linear function approximati
on and policy evaluation setting, we consider this as a meaningful step towards
the theoretical understanding of deep Q-learning variants with target networks.
****
Functional Transparency for Structured Data: a Game-Theoretic Approach
Guang-He Lee, Wengong Jin, David Alvarez-Melis, Tommi Jaakkola
We provide a new approach to training neural models to exhibit transparency in a
 well-defined, functional manner. Our approach naturally operates over structure
d data and tailors the predictor, functionally, towards a chosen family of (loca
l) witnesses. The estimation problem is setup as a co-operative game between an
unrestricted predictor such as a neural network, and a set of witnesses chosen f
rom the desired transparent family. The goal of the witnesses is to highlight, l
ocally, how well the predictor conforms to the chosen family of functions, while
 the predictor is trained to minimize the highlighted discrepancy. We emphasize
that the predictor remains globally powerful as it is only encouraged to agree l
ocally with locally adapted witnesses. We analyze the effect of the proposed app
roach, provide example formulations in the context of deep graph and sequence mo
dels, and empirically illustrate the idea in chemical property prediction, tempo
ral modeling, and molecule representation learning.
****
Self-Attention Graph Pooling
Junhyun Lee, Inyeop Lee, Jaewoo Kang
Advanced methods of applying deep learning to structured data such as graphs hav
e been proposed in recent years. In particular, studies have focused on generali
zing convolutional neural networks to graph data, which includes redefining the
convolution and the downsampling (pooling) operations for graphs. The method of
generalizing the convolution operation to graphs has been proven to improve perf
ormance and is widely used. However, the method of applying downsampling to grap
hs is still difficult to perform and has room for improvement. In this paper, we
 propose a graph pooling method based on self-attention. Self-attention using gr
aph convolution allows our pooling method to consider both node features and gra
ph topology. To ensure a fair comparison, the same training procedures and model
 architectures were used for the existing pooling methods and our method. The ex
perimental results demonstrate that our method achieves superior graph classific
ation performance on the benchmark datasets using a reasonable number of paramet
ers.
****
Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Ne
tworks
Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, Yee Whye Teh
Many machine learning tasks such as multiple instance learning, 3D shape recogni
tion, and few-shot image classification are defined on sets of instances. Since
solutions to such problems do not depend on the order of elements of the set, mo
dels used to address them should be permutation invariant. We present an attenti
on-based neural network module, the Set Transformer, specifically designed to mo
del interactions among elements in the input set. The model consists of an encod
er and a decoder, both of which rely on attention mechanisms. In an effort to re
duce computational complexity, we introduce an attention scheme inspired by indu
cing point methods from sparse Gaussian process literature. It reduces the compu
tation time of self-attention from quadratic to linear in the number of elements
 in the set. We show that our model is theoretically attractive and we evaluate
it on a range of tasks, demonstrating the state-of-the-art performance compared

to recent methods for set-structured data.
****
First-Order Algorithms Converge Faster than $O(1/k)$ on Convex Problems
Ching-Pei Lee, Stephen Wright
It is well known that both gradient descent and stochastic coordinate descent achieve a global convergence rate of $O(1/k)$ in the objective value, when applied to a scheme for minimizing a Lipschitz-continuously differentiable, unconstrained convex function. In this work, we improve this rate to $o(1/k)$. We extend the result to proximal gradient and proximal coordinate descent on regularized problems to show similar $o(1/k)$ convergence rates. The result is tight in the sense that a rate of $O(1/k^{1+\epsilon})$ is not generally attainable for any $\epsilon>0$, for any of these methods.
****
Robust Inference via Generative Classifiers for Handling Noisy Labels
Kimin Lee, Sukmin Yun, Kibok Lee, Honglak Lee, Bo Li, Jinwoo Shin
Large-scale datasets may contain significant proportions of noisy (incorrect) class labels, and it is well-known that modern deep neural networks (DNNs) poorly generalize from such noisy training datasets. To mitigate the issue, we propose a novel inference method, termed Robust Generative classifier (RoG), applicable to any discriminative (e.g., softmax) neural classifier pre-trained on noisy datasets. In particular, we induce a generative classifier on top of hidden feature spaces of the pre-trained DNNs, for obtaining a more robust decision boundary. By estimating the parameters of generative classifier using the minimum covariance determinant estimator, we significantly improve the classification accuracy with neither re-training of the deep model nor changing its architectures. With the assumption of Gaussian distribution for features, we prove that RoG generalizes better than baselines under noisy labels. Finally, we propose the ensemble version of RoG to improve its performance by investigating the layer-wise characteristics of DNNs. Our extensive experimental results demonstrate the superiority of RoG given different learning models optimized by several training techniques to handle diverse scenarios of noisy labels.
****
Sublinear Time Nearest Neighbor Search over Generalized Weighted Space
Yifan Lei, Qiang Huang, Mohan Kankanhalli, Anthony Tung
Nearest Neighbor Search (NNS) over generalized weighted space is a fundamental problem which has many applications in various fields. However, to the best of our knowledge, there is no sublinear time solution to this problem. Based on the idea of Asymmetric Locality-Sensitive Hashing (ALSH), we introduce a novel spherical asymmetric transformation and propose the first two novel weight-oblivious hashing schemes SL-ALSH and S2-ALSH accordingly. We further show that both schemes enjoy a quality guarantee and can answer the NNS queries in sublinear time. Evaluations over three real datasets demonstrate the superior performance of the two proposed schemes.
****
MONK Outlier-Robust Mean Embedding Estimation by Median-of-Means
Matthieu Lerasle, Zoltan Szabo, Timothée Mathieu, Guillaume Lecue
Mean embeddings provide an extremely flexible and powerful tool in machine learning and statistics to represent probability distributions and define a semi-metric (MMD, maximum mean discrepancy; also called N-distance or energy distance), with numerous successful applications. The representation is constructed as the expectation of the feature map defined by a kernel. As a mean, its classical empirical estimator, however, can be arbitrary severely affected even by a single outlier in case of unbounded features. To the best of our knowledge, unfortunately even the consistency of the existing few techniques trying to alleviate this serious sensitivity bottleneck is unknown. In this paper, we show how the recently emerged principle of median-of-means can be used to design estimators for kernel mean embedding and MMD with excessive resistance properties to outliers, and optimal sub-Gaussian deviation bounds under mild assumptions.
****
Cheap Orthogonal Constraints in Neural Networks: A Simple Parametrization of the

Orthogonal and Unitary Group
Mario Lezcano-Casado, David Mart■■nez-Rubio
We introduce a novel approach to perform first-order optimization with orthogonal and unitary constraints. This approach is based on a parametrization stemming from Lie group theory through the exponential map. The parametrization transforms the constrained optimization problem into an unconstrained one over a Euclidean space, for which common first-order optimization methods can be used. The theoretical results presented are general enough to cover the special orthogonal group, the unitary group and, in general, any connected compact Lie group. We discuss how this and other parametrizations can be computed efficiently through an implementation trick, making numerically complex parametrizations usable at a negligible runtime cost in neural networks. In particular, we apply our results to RNNs with orthogonal recurrent weights, yielding a new architecture called expRNN. We demonstrate how our method constitutes a more robust approach to optimization with orthogonal constraints, showing faster, accurate, and more stable convergence in several tasks designed to test RNNs.
****
Are Generative Classifiers More Robust to Adversarial Attacks?
Yingzhen Li, John Bradshaw, Yash Sharma
There is a rising interest in studying the robustness of deep neural network classifiers against adversaries, with both advanced attack and defence techniques being actively developed. However, most recent work focuses on discriminative classifiers, which only model the conditional distribution of the labels given the inputs. In this paper, we propose and investigate the deep Bayes classifier, which improves classical naive Bayes with conditional deep generative models. We further develop detection methods for adversarial examples, which reject inputs with low likelihood under the generative model. Experimental results suggest that deep Bayes classifiers are more robust than deep discriminative classifiers, and that the proposed detection methods are effective against many recently proposed attacks.
****
Sublinear quantum algorithms for training linear and kernel-based classifiers
Tongyang Li, Shouvanik Chakrabarti, Xiaodi Wu
We investigate quantum algorithms for classification, a fundamental problem in machine learning, with provable guarantees. Given n d-dimensional data points, the state-of-the-art (and optimal) classical algorithm for training classifiers with constant margin by Clarkson et al. runs in $\tilde{O}(n +d)$, which is also optimal in its input/output model. We design sublinear quantum algorithms for the same task running in $\tilde{O}(\sqrt{n} +\sqrt{d})$, a quadratic improvement in both n and d. Moreover, our algorithms use the standard quantization of the classical input and generate the same classical output, suggesting minimal overheads when used as subroutines for end-to-end applications. We also demonstrate a tight lower bound (up to poly-log factors) and discuss the possibility of implementation on near-term quantum machines.
****
LGM-Net: Learning to Generate Matching Networks for Few-Shot Learning
Huaiyu Li, Weiming Dong, Xing Mei, Chongyang Ma, Feiyue Huang, Bao-Gang Hu
In this work, we propose a novel meta-learning approach for few-shot classification, which learns transferable prior knowledge across tasks and directly produces network parameters for similar unseen tasks with training samples. Our approach, called LGM-Net, includes two key modules, namely, TargetNet and MetaNet. The TargetNet module is a neural network for solving a specific task and the MetaNet module aims at learning to generate functional weights for TargetNet by observing training samples. We also present an intertask normalization strategy for the training process to leverage common information shared across different tasks. The experimental results on Omniglot and miniImageNet datasets demonstrate that LGM-Net can effectively adapt to similar unseen tasks and achieve competitive performance, and the results on synthetic datasets show that transferable prior knowledge is learned by the MetaNet module via mapping training data to functional weights. LGM-Net enables fast learning and adaptation since no further tuning s

teps are required compared to other meta-learning approaches
****
Graph Matching Networks for Learning the Similarity of Graph Structured Objects
Yujia Li, Chenjie Gu, Thomas Dullien, Oriol Vinyals, Pushmeet Kohli
This paper addresses the challenging problem of retrieval and matching of graph structured objects, and makes two key contributions. First, we demonstrate how Graph Neural Networks (GNN), which have emerged as an effective model for various supervised prediction problems defined on structured data, can be trained to produce embedding of graphs in vector spaces that enables efficient similarity reasoning. Second, we propose a novel Graph Matching Network model that, given a pair of graphs as input, computes a similarity score between them by jointly reasoning on the pair through a new cross-graph attention-based matching mechanism. We demonstrate the effectiveness of our models on different domains including the challenging problem of control-flow graph based function similarity search that plays an important role in the detection of vulnerabilities in software systems. The experimental analysis demonstrates that our models are not only able to exploit structure in the context of similarity learning but they can also outperform domain specific baseline systems that have been carefully hand-engineered for these problems.
****
Area Attention
Yang Li, Lukasz Kaiser, Samy Bengio, Si Si
Existing attention mechanisms are trained to attend to individual items in a collection (the memory) with a predefined, fixed granularity, e.g., a word token or an image grid. We propose area attention: a way to attend to areas in the memory, where each area contains a group of items that are structurally adjacent, e.g., spatially for a 2D memory such as images, or temporally for a 1D memory such as natural language sentences. Importantly, the shape and the size of an area are dynamically determined via learning, which enables a model to attend to information with varying granularity. Area attention can easily work with existing model architectures such as multi-head attention for simultaneously attending to multiple areas in the memory. We evaluate area attention on two tasks: neural machine translation (both character and token-level) and image captioning, and improve upon strong (state-of-the-art) baselines in all the cases. These improvements are obtainable with a basic form of area attention that is parameter free.
****
Online Learning to Rank with Features
Shuai Li, Tor Lattimore, Csaba Szepesvari
We introduce a new model for online ranking in which the click probability factors into an examination and attractiveness function and the attractiveness function is a linear function of a feature vector and an unknown parameter. Only relatively mild assumptions are made on the examination function. A novel algorithm for this setup is analysed, showing that the dependence on the number of items is replaced by a dependence on the dimension, allowing the new algorithm to handle a large number of items. When reduced to the orthogonal case, the regret of the algorithm improves on the state-of-the-art.
****
NATTACK: Learning the Distributions of Adversarial Examples for an Improved Black-Box Attack on Deep Neural Networks
Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, Boqing Gong
Powerful adversarial attack methods are vital for understanding how to construct robust deep neural networks (DNNs) and for thoroughly testing defense techniques. In this paper, we propose a black-box adversarial attack algorithm that can defeat both vanilla DNNs and those generated by various defense techniques developed recently. Instead of searching for an "optimal" adversarial example for a benign input to a targeted DNN, our algorithm finds a probability density distribution over a small region centered around the input, such that a sample drawn from this distribution is likely an adversarial example, without the need of accessing the DNN’s internal layers or weights. Our approach is universal as it can successfully attack different neural networks by a single algorithm. It is also st

rong; according to the testing against 2 vanilla DNNs and 13 defended ones, it outperforms state-of-the-art black-box or white-box attack methods for most test cases. Additionally, our results reveal that adversarial training remains one of the best defense techniques, and the adversarial examples are not as transferable across defended DNNs as them across vanilla DNNs.
****
Bayesian Joint Spike-and-Slab Graphical Lasso
Zehang Li, Tyler Mccormick, Samuel Clark
In this article, we propose a new class of priors for Bayesian inference with multiple Gaussian graphical models. We introduce Bayesian treatments of two popular procedures, the group graphical lasso and the fused graphical lasso, and extend them to a continuous spike-and-slab framework to allow self-adaptive shrinkage and model selection simultaneously. We develop an EM algorithm that performs fast and dynamic explorations of posterior modes. Our approach selects sparse models efficiently and automatically with substantially smaller bias than would be induced by alternative regularization procedures. The performance of the proposed methods are demonstrated through simulation and two real data examples.
****
Exploiting Worker Correlation for Label Aggregation in Crowdsourcing
Yuan Li, Benjamin Rubinstein, Trevor Cohn
Crowdsourcing has emerged as a core component of data science pipelines. From collected noisy worker labels, aggregation models that incorporate worker reliability parameters aim to infer a latent true annotation. In this paper, we argue that existing crowdsourcing approaches do not sufficiently model worker correlations observed in practical settings; we propose in response an enhanced Bayesian classifier combination (EBCC) model, with inference based on a mean-field variational approach. An introduced mixture of intra-class reliabilities—connected to tensor decomposition and item clustering—induces inter-worker correlation. EBCC does not suffer the limitations of existing correlation models: intractable marginalisation of missing labels and poor scaling to large worker cohorts. Extensive empirical comparison on 17 real-world datasets sees EBCC achieving the highest mean accuracy across 10 benchmark crowdsourcing methods.
****
Adversarial camera stickers: A physical camera-based attack on deep learning systems
Juncheng Li, Frank Schmidt, Zico Kolter
Recent work has documented the susceptibility of deep learning systems to adversarial examples, but most such attacks directly manipulate the digital input to a classifier. Although a smaller line of work considers physical adversarial attacks, in all cases these involve manipulating the object of interest, e.g., putting a physical sticker on an object to misclassify it, or manufacturing an object specifically intended to be misclassified. In this work, we consider an alternative question: is it possible to fool deep classifiers, over all perceived objects of a certain type, by physically manipulating the camera itself? We show that by placing a carefully crafted and mainly-translucent sticker over the lens of a camera, one can create universal perturbations of the observed images that are inconspicuous, yet misclassify target objects as a different (targeted) class. To accomplish this, we propose an iterative procedure for both updating the attack perturbation (to make it adversarial for a given classifier), and the threat model itself (to ensure it is physically realizable). For example, we show that we can achieve physically-realizable attacks that fool ImageNet classifiers in a targeted fashion 49.6% of the time. This presents a new class of physically-realizable threat models to consider in the context of adversarially robust machine learning. Our demo video can be viewed at: https://youtu.be/wUVmL33Fx54
****
Towards a Unified Analysis of Random Fourier Features
Zhu Li, Jean-Francois Ton, Dino Oglic, Dino Sejdinovic
Random Fourier features is a widely used, simple, and effective technique for scaling up kernel methods. The existing theoretical analysis of the approach, however, remains focused on specific learning tasks and typically gives pessimistic

bounds which are at odds with the empirical results. We tackle these problems and provide the first unified risk analysis of learning with random Fourier features using the squared error and Lipschitz continuous loss functions. In our bounds, the trade-off between the computational cost and the expected risk convergence rate is problem specific and expressed in terms of the regularization parameter and the number of effective degrees of freedom. We study both the standard random Fourier features method for which we improve the existing bounds on the number of features required to guarantee the corresponding minimax risk convergence rate of kernel ridge regression, as well as a data-dependent modification which samples features proportional to ridge leverage scores and further reduces the required number of features. As ridge leverage scores are expensive to compute, we devise a simple approximation scheme which provably reduces the computational cost without loss of statistical efficiency.
****
Feature-Critic Networks for Heterogeneous Domain Generalization
Yiying Li, Yongxin Yang, Wei Zhou, Timothy Hospedales
The well known domain shift issue causes model performance to degrade when deployed to a new target domain with different statistics to training. Domain adaptation techniques alleviate this, but need some instances from the target domain to drive adaptation. Domain generalisation is the recently topical problem of learning a model that generalises to unseen domains out of the box, and various approaches aim to train a domain-invariant feature extractor, typically by adding some manually designed losses. In this work, we propose a learning to learn approach, where the auxiliary loss that helps generalisation is itself learned. Beyond conventional domain generalisation, we consider a more challenging setting of heterogeneous domain generalisation, where the unseen domains do not share label space with the seen ones, and the goal is to train a feature representation that is useful off-the-shelf for novel data and novel categories. Experimental evaluation demonstrates that our method outperforms state-of-the-art solutions in both settings.
****
Learn to Grow: A Continual Structure Learning Framework for Overcoming Catastrophic Forgetting
Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, Caiming Xiong
Addressing catastrophic forgetting is one of the key challenges in continual learning where machine learning systems are trained with sequential or streaming tasks. Despite recent remarkable progress in state-of-the-art deep learning, deep neural networks (DNNs) are still plagued with the catastrophic forgetting problem. This paper presents a conceptually simple yet general and effective framework for handling catastrophic forgetting in continual learning with DNNs. The proposed method consists of two components: a neural structure optimization component and a parameter learning and/or fine-tuning component. By separating the explicit neural structure learning and the parameter estimation, not only is the proposed method capable of evolving neural structures in an intuitively meaningful way, but also shows strong capabilities of alleviating catastrophic forgetting in experiments. Furthermore, the proposed method outperforms all other baselines on the permuted MNIST dataset, the split CIFAR100 dataset and the Visual Domain Decathlon dataset in continual learning setting.
****
Alternating Minimizations Converge to Second-Order Optimal Solutions
Qiuwei Li, Zhihui Zhu, Gongguo Tang
This work studies the second-order convergence for both standard alternating minimization and proximal alternating minimization. We show that under mild assumptions on the (nonconvex) objective function, both algorithms avoid strict saddles almost surely from random initialization. Together with known first-order convergence results, this implies both algorithms converge to a second-order stationary point. This solves an open problem for the second-order convergence of alternating minimization algorithms that have been widely used in practice to solve large-scale nonconvex problems due to their simple implementation, fast convergence, and superb empirical performance.

****
Cautious Regret Minimization: Online Optimization with Long-Term Budget Constraints
Nikolaos Liakopoulos, Apostolos Destounis, Georgios Paschos, Thrasyvoulos Spyropoulos, Panayotis Mertikopoulos
We study a class of online convex optimization problems with long-term budget constraints that arise naturally as reliability guarantees or total consumption constraints. In this general setting, prior work by Mannor et al. (2009) has shown that achieving no regret is impossible if the functions defining the agent’s budget are chosen by an adversary. To overcome this obstacle, we refine the agent’s regret metric by introducing the notion of a "K-benchmark", i.e., a comparator which meets the problem’s allotted budget over any window of length K. The impossibility analysis of Mannor et al. (2009) is recovered when K=T; however, for K=o(T), we show that it is possible to minimize regret while still meeting the problem’s long-term budget constraints. We achieve this via an online learning policy based on Cautious Online Lagrangiant Descent (COLD) for which we derive explicit bounds, in terms of both the incurred regret and the residual budget violations.
****
Regularization in directable environments with application to Tetris
Jan Malte Lichtenberg, Özgür ■im■ek
Learning from small data sets is difficult in the absence of specific domain knowledge. We present a regularized linear model called STEW that benefits from a generic and prevalent form of prior knowledge: feature directions. STEW shrinks weights toward each other, converging to an equal-weights solution in the limit of infinite regularization. We provide theoretical results on the equal-weights solution that explains how STEW can productively trade-off bias and variance. Across a wide range of learning problems, including Tetris, STEW outperformed existing linear models, including ridge regression, the Lasso, and the non-negative Lasso, when feature directions were known. The model proved to be robust to unreliable (or absent) feature directions, still outperforming alternative models under diverse conditions. Our results in Tetris were obtained by using a novel approach to learning in sequential decision environments based on multinomial logistic regression.
****
Inference and Sampling of K_33-free Ising Models
Valerii Likhosherstov, Yury Maximov, Misha Chertkov
We call an Ising model tractable when it is possible to compute its partition function value (statistical inference) in polynomial time. The tractability also implies an ability to sample configurations of this model in polynomial time. The notion of tractability extends the basic case of planar zero-field Ising models. Our starting point is to describe algorithms for the basic case, computing partition function and sampling efficiently. Then, we extend our tractable inference and sampling algorithms to models whose triconnected components are either planar or graphs of $O(1)$ size. In particular, it results in a polynomial-time inference and sampling algorithms for K_{33} (minor)-free topologies of zero-field Ising models—a generalization of planar graphs with a potentially unbounded genus.
****
Kernel-Based Reinforcement Learning in Robust Markov Decision Processes
Shiau Hong Lim, Arnaud Autef
The robust Markov decision processes (MDP) framework aims to address the problem of parameter uncertainty due to model mismatch, approximation errors or even adversarial behaviors. It is especially relevant when deploying the learned policies in real-world applications. Scaling up the robust MDP framework to large or continuous state space remains a challenging problem. The use of function approximation in this case is usually inevitable and this can only amplify the problem of model mismatch and parameter uncertainties. It has been previously shown that, in the case of MDPs with state aggregation, the robust policies enjoy a tighter performance bound compared to standard solutions due to its reduced sensitivit

y to approximation errors. We extend these results to the much larger class of kernel-based approximators and show, both analytically and empirically that the robust policies can significantly outperform the non-robust counterpart.
****
On Efficient Optimal Transport: An Analysis of Greedy and Accelerated Mirror Descent Algorithms
Tianyi Lin, Nhat Ho, Michael Jordan
We provide theoretical analyses for two algorithms that solve the regularized optimal transport (OT) problem between two discrete probability measures with at most n atoms. We show that a greedy variant of the classical Sinkhorn algorithm, known as the Greenkhorn algorithm, can be improved to $\bigOtil\left(n^2/\varepsilon^2\right)$, improving on the best known complexity bound of $\bigOtil\left(n^2/\varepsilon^3\right)$. This matches the best known complexity bound for the Sinkhorn algorithm and helps explain why the Greenkhorn algorithm outperforms the Sinkhorn algorithm in practice. Our proof technique is based on a primal-dual formulation and provide a tight upper bound for the dual solution, leading to a class of adaptive primal-dual accelerated mirror descent (APDAMD) algorithms. We prove that the complexity of these algorithms is $\bigOtil\left(n^2\sqrt{\gamma}/\varepsilon\right)$ in which $\gamma \in (0, n]$ refers to some constants in the Bregman divergence. Experimental results on synthetic and real datasets demonstrate the favorable performance of the Greenkhorn and APDAMD algorithms in practice.
****
Fast and Simple Natural-Gradient Variational Inference with Mixture of Exponential-family Approximations
Wu Lin, Mohammad Emtiyaz Khan, Mark Schmidt
Natural-gradient methods enable fast and simple algorithms for variational inference, but due to computational difficulties, their use is mostly limited to minimal exponential-family (EF) approximations. In this paper, we extend their application to estimate structured approximations such as mixtures of EF distributions. Such approximations can fit complex, multimodal posterior distributions and are generally more accurate than unimodal EF approximations. By using a minimal conditional-EF representation of such approximations, we derive simple natural-gradient updates. Our empirical results demonstrate a faster convergence of our natural-gradient method compared to black-box gradient-based methods. Our work expands the scope of natural gradients for Bayesian inference and makes them more widely applicable than before.
****
Acceleration of SVRG and Katyusha X by Inexact Preconditioning
Yanli Liu, Fei Feng, Wotao Yin
Empirical risk minimization is an important class of optimization problems with many popular machine learning applications, and stochastic variance reduction methods are popular choices for solving them. Among these methods, SVRG and Katyusha X (a Nesterov accelerated SVRG) achieve fast convergence without substantial memory requirement. In this paper, we propose to accelerate these two algorithms by inexact preconditioning, the proposed methods employ fixed preconditioners, although the subproblem in each epoch becomes harder, it suffices to apply fixed number of simple subroutines to solve it inexactly, without losing the overall convergence. As a result, this inexact preconditioning strategy gives provably better iteration complexity and gradient complexity over SVRG and Katyusha X. We also allow each function in the finite sum to be nonconvex while the sum is strongly convex. In our numerical experiments, we observe an on average $8\times$ speedup on the number of iterations and $7\times$ speedup on runtime.
****
Transferable Adversarial Training: A General Approach to Adapting Deep Classifiers
Hong Liu, Mingsheng Long, Jianmin Wang, Michael Jordan
Domain adaptation enables knowledge transfer from a labeled source domain to an unlabeled target domain. A mainstream approach is adversarial feature adaptation, which learns domain-invariant representations through aligning the feature dis

tributions of both domains. However, a theoretical prerequisite of domain adapta tion is the adaptability measured by the expected risk of an ideal joint hypothe sis over the source and target domains. In this respect, adversarial feature ada ptation may potentially deteriorate the adaptability, since it distorts the orig inal feature distributions when suppressing domain-specific variations. To this end, we propose Transferable Adversarial Training (TAT) to enable the adaptation of deep classifiers. The approach generates transferable examples to fill in th e gap between the source and target domains, and adversarially trains the deep c lassifiers to make consistent predictions over the transferable examples. Withou t learning domain-invariant representations at the expense of distorting the fea ture distributions, the adaptability in the theoretical learning bound is algori thmically guaranteed. A series of experiments validate that our approach advance s the state of the arts on a variety of domain adaptation tasks in vision and NL P, including object recognition, learning from synthetic to real data, and senti ment classification.
****
Rao-Blackwellized Stochastic Gradients for Discrete Distributions
Runjing Liu, Jeffrey Regier, Nilesh Tripuraneni, Michael Jordan, Jon Mcauliffe
We wish to compute the gradient of an expectation over a finite or countably inf inite sample space having K $\leq$ ∞ categories. When K is indeed infinit e, or finite but very large, the relevant summation is intractable. Accordingly, various stochastic gradient estimators have been proposed. In this paper, we de scribe a technique that can be applied to reduce the variance of any such estima tor, without changing its bias{—}in particular, unbiasedness is retained. We sho w that our technique is an instance of Rao-Blackwellization, and we demonstrate the improvement it yields on a semi-supervised classification problem and a pixe l attention task.
****
Sparse Extreme Multi-label Learning with Oracle Property
Weiwei Liu, Xiaobo Shen
The pioneering work of sparse local embeddings for extreme classification (SLEEC) (Bhatia et al., 2015) has shown great promise in multi-label learning. Unfortu nately, the statistical rate of convergence and oracle property of SLEEC are sti ll not well understood. To fill this gap, we present a unified framework for SLE EC with nonconvex penalty. Theoretically, we rigorously prove that our proposed estimator enjoys oracle property (i.e., performs as well as if the underlying mo del were known beforehand), and obtains a desirable statistical convergence rate . Moreover, we show that under a mild condition on the magnitude of the entries in the underlying model, we are able to obtain an improved convergence rate. Ext ensive numerical experiments verify our theoretical findings and the superiority of our proposed estimator.
****
Data Poisoning Attacks on Stochastic Bandits
Fang Liu, Ness Shroff
Stochastic multi-armed bandits form a class of online learning problems that hav e important applications in online recommendation systems, adaptive medical trea tment, and many others. Even though potential attacks against these learning alg orithms may hijack their behavior, causing catastrophic loss in real-world appli cations, little is known about adversarial attacks on bandit algorithms. In this paper, we propose a framework of offline attacks on bandit algorithms and study convex optimization based attacks on several popular bandit algorithms. We show that the attacker can force the bandit algorithm to pull a target arm with high probability by a slight manipulation of the rewards in the data. Then we study a form of online attacks on bandit algorithms and propose an adaptive attack str ategy against any bandit algorithm without the knowledge of the bandit algorithm . Our adaptive attack strategy can hijack the behavior of the bandit algorithm t o suffer a linear regret with only a logarithmic cost to the attacker. Our resul ts demonstrate a significant security threat to stochastic bandits.
****
The Implicit Fairness Criterion of Unconstrained Learning

Lydia T. Liu, Max Simchowitz, Moritz Hardt

We clarify what fairness guarantees we can and cannot expect to follow from unconstrained machine learning. Specifically, we show that in many settings, unconstrained learning on its own implies group calibration, that is, the outcome variable is conditionally independent of group membership given the score. A lower bound confirms the optimality of our upper bound. Moreover, we prove that as the excess risk of the learned score decreases, the more strongly it violates separation and independence, two other standard fairness criteria. Our results challenge the view that group calibration necessitates an active intervention, suggesting that often we ought to think of it as a byproduct of unconstrained machine learning.

****

Taming MAML: Efficient unbiased meta-reinforcement learning

Hao Liu, Richard Socher, Caiming Xiong

While meta reinforcement learning (Meta-RL) methods have achieved remarkable success, obtaining correct and low variance estimates for policy gradients remains a significant challenge. In particular, estimating a large Hessian, poor sample efficiency and unstable training continue to make Meta-RL difficult. We propose a surrogate objective function named, Taming MAML (TMAML), that adds control variates into gradient estimation via automatic differentiation. TMAML improves the quality of gradient estimation by reducing variance without introducing bias. We further propose a version of our method that extends the meta-learning framework to learning the control variates themselves, enabling efficient and scalable learning from a distribution of MDPs. We empirically compare our approach with MAML and other variance-bias trade-off methods including DICE, LVC, and action-dependent control variates. Our approach is easy to implement and outperforms existing methods in terms of the variance and accuracy of gradient estimation, ultimately yielding higher performance across a variety of challenging Meta-RL environments.

****

On Certifying Non-Uniform Bounds against Adversarial Attacks

Chen Liu, Ryota Tomioka, Volkan Cevher

This work studies the robustness certification problem of neural network models, which aims to find certified adversary-free regions as large as possible around data points. In contrast to the existing approaches that seek regions bounded uniformly along all input features, we consider non-uniform bounds and use it to study the decision boundary of neural network models. We formulate our target as an optimization problem with nonlinear constraints. Then, a framework applicable for general feedforward neural networks is proposed to bound the output logits so that the relaxed problem can be solved by the augmented Lagrangian method. Our experiments show the non-uniform bounds have larger volumes than uniform ones. Compared with normal models, the robust models have even larger non-uniform bounds and better interpretability. Further, the geometric similarity of the non-uniform bounds gives a quantitative, data-agnostic metric of input features’ robustness.

****

Understanding and Accelerating Particle-Based Variational Inference

Chang Liu, Jingwei Zhuo, Pengyu Cheng, Ruiyi Zhang, Jun Zhu

Particle-based variational inference methods (ParVIs) have gained attention in the Bayesian inference literature, for their capacity to yield flexible and accurate approximations. We explore ParVIs from the perspective of Wasserstein gradient flows, and make both theoretical and practical contributions. We unify various finite-particle approximations that existing ParVIs use, and recognize that the approximation is essentially a compulsory smoothing treatment, in either of two equivalent forms. This novel understanding reveals the assumptions and relations of existing ParVIs, and also inspires new ParVIs. We propose an acceleration framework and a principled bandwidth-selection method for general ParVIs; these are based on the developed theory and leverage the geometry of the Wasserstein space. Experimental results show the improved convergence by the acceleration framework and enhanced sample accuracy by the bandwidth-selection method.

****
Understanding MCMC Dynamics as Flows on the Wasserstein Space
Chang Liu, Jingwei Zhuo, Jun Zhu
It is known that the Langevin dynamics used in MCMC is the gradient flow of the
KL divergence on the Wasserstein space, which helps convergence analysis and ins
pires recent particle-based variational inference methods (ParVIs). But no more
MCMC dynamics is understood in this way. In this work, by developing novel conce
pts, we propose a theoretical framework that recognizes a general MCMC dynamics
as the fiber-gradient Hamiltonian flow on the Wasserstein space of a fiber-Riema
nnian Poisson manifold. The "conservation + convergence" structure of the flow g
ives a clear picture on the behavior of general MCMC dynamics. The framework als
o enables ParVI simulation of MCMC dynamics, which enriches the ParVI family wit
h more efficient dynamics, and also adapts ParVI advantages to MCMCs. We develop
 two ParVI methods for a particular MCMC dynamics and demonstrate the benefits i
n experiments.
****
Sliced-Wasserstein Flows: Nonparametric Generative Modeling via Optimal Transpor
t and Diffusions
Antoine Liutkus, Umut Simsekli, Szymon Majewski, Alain Durmus, Fabian-Robert Stö
ter
By building upon the recent theory that established the connection between impli
cit generative modeling (IGM) and optimal transport, in this study, we propose a
 novel parameter-free algorithm for learning the underlying distributions of com
plicated datasets and sampling from them. The proposed algorithm is based on a f
unctional optimization problem, which aims at finding a measure that is close to
 the data distribution as much as possible and also expressive enough for genera
tive modeling purposes. We formulate the problem as a gradient flow in the space
 of probability measures. The connections between gradient flows and stochastic
differential equations let us develop a computationally efficient algorithm for
solving the optimization problem. We provide formal theoretical analysis where w
e prove finite-time error guarantees for the proposed algorithm. To the best of
our knowledge, the proposed algorithm is the first nonparametric IGM algorithm w
ith explicit theoretical guarantees. Our experimental results support our theory
 and show that our algorithm is able to successfully capture the structure of di
fferent types of data distributions.
****
Challenging Common Assumptions in the Unsupervised Learning of Disentangled Repr
esentations
Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, B
ernhard Schölkopf, Olivier Bachem
The key idea behind the unsupervised learning of disentangled representations is
 that real-world data is generated by a few explanatory factors of variation whi
ch can be recovered by unsupervised learning algorithms. In this paper, we provi
de a sober look at recent progress in the field and challenge some common assump
tions. We first theoretically show that the unsupervised learning of disentangle
d representations is fundamentally impossible without inductive biases on both t
he models and the data. Then, we train more than 12000 models covering most pr
ominent methods and evaluation metrics in a reproducible large-scale experimenta
l study on seven different data sets. We observe that while the different method
s successfully enforce properties “encouraged” by the corresponding losses, well
-disentangled models seemingly cannot be identified without supervision. Further
more, increased disentanglement does not seem to lead to a decreased sample comp
lexity of learning for downstream tasks. Our results suggest that future work on
 disentanglement learning should be explicit about the role of inductive biases
and (implicit) supervision, investigate concrete benefits of enforcing disentang
lement of the learned representations, and consider a reproducible experimental
setup covering several data sets.
****
Bayesian Counterfactual Risk Minimization
Ben London, Ted Sandler

We present a Bayesian view of counterfactual risk minimization (CRM) for offline
 learning from logged bandit feedback. Using PAC-Bayesian analysis, we derive a
new generalization bound for the truncated inverse propensity score estimator. W
e apply the bound to a class of Bayesian policies, which motivates a novel, pote
ntially data-dependent, regularization technique for CRM. Experimental results i
ndicate that this technique outperforms standard L_2 regularization, and that
it is competitive with variance regularization while being both simpler to imple
ment and more computationally efficient.
****
PA-GD: On the Convergence of Perturbed Alternating Gradient Descent to Second-Or
der Stationary Points for Structured Nonconvex Optimization
Songtao Lu, Mingyi Hong, Zhengdao Wang
Alternating gradient descent (A-GD) is a simple but popular algorithm in machine
 learning, which updates two blocks of variables in an alternating manner using
gradient descent steps. In this paper, we consider a smooth unconstrained noncon
vex optimization problem, and propose a perturbed A-GD (PA-GD) which is able to
converge (with high probability) to the second-order stationary points (SOSPs) w
ith a global sublinear rate. Existing analysis on A-GD type algorithm either onl
y guarantees convergence to first-order solutions, or converges to second-order
solutions asymptotically (without rates). To the best of our knowledge, this is
the first alternating type algorithm that takes $\mathcal{O}(\text{polylog}(d)/\
epsilon^2)$ iterations to achieve an ($\epsilon,\sqrt{\epsilon}$)-SOSP with high
 probability, where polylog(d) denotes the polynomial of the logarithm with re
spect to problem dimension d.
****
Neurally-Guided Structure Inference
Sidi Lu, Jiayuan Mao, Joshua Tenenbaum, Jiajun Wu
Most structure inference methods either rely on exhaustive search or are purely
data-driven. Exhaustive search robustly infers the structure of arbitrarily comp
lex data, but it is slow. Data-driven methods allow efficient inference, but do
not generalize when test data have more complex structures than training data. I
n this paper, we propose a hybrid inference algorithm, the Neurally-Guided Struc
ture Inference (NG-SI), keeping the advantages of both search-based and data-dri
ven methods. The key idea of NG-SI is to use a neural network to guide the hiera
rchical, layer-wise search over the compositional space of structures. We evalua
te our algorithm on two representative structure inference tasks: probabilistic
matrix decomposition and symbolic program parsing. It outperforms data-driven an
d search-based alternatives on both tasks.
****
Optimal Algorithms for Lipschitz Bandits with Heavy-tailed Rewards
Shiyin Lu, Guanghui Wang, Yao Hu, Lijun Zhang
We study Lipschitz bandits, where a learner repeatedly plays one arm from an inf
inite arm set and then receives a stochastic reward whose expectation is a Lipsc
hitz function of the chosen arm. Most of existing work assume the reward distrib
utions are bounded or at least sub-Gaussian, and thus do not apply to heavy-tail
ed rewards arising in many real-world scenarios such as web advertising and fina
ncial markets. To address this limitation, in this paper we relax the assumption
 on rewards to allow arbitrary distributions that have finite $(1+\epsilon)$-th
moments for some $\epsilon \in (0, 1]$, and propose algorithms that enjoy a subl
inear regret of $\widetilde{O}(T^{(d_z\epsilon + 1)/(d_z \epsilon + \epsilon + 1
)})$ where T is the time horizon and d_z is the zooming dimension. The key i
dea is to exploit the Lipschitz property of the expected reward function by adap
tively discretizing the arm set, and employ upper confidence bound policies with
 robust mean estimators designed for heavy-tailed distributions. Furthermore, we
 provide a lower bound for Lipschitz bandits with heavy-tailed rewards, and show
 that our algorithms are optimal in terms of T. Finally, we conduct numerical
experiments to demonstrate the effectiveness of our algorithms.
****
CoT: Cooperative Training for Generative Modeling of Discrete Data
Sidi Lu, Lantao Yu, Siyuan Feng, Yaoming Zhu, Weinan Zhang

In this paper, we study the generative models of sequential discrete data. To tackle the exposure bias problem inherent in maximum likelihood estimation (MLE), generative adversarial networks (GANs) are introduced to penalize the unrealistic generated samples. To exploit the supervision signal from the discriminator, most previous models leverage REINFORCE to address the non-differentiable problem of sequential discrete data. However, because of the unstable property of the training signal during the dynamic process of adversarial training, the effectiveness of REINFORCE, in this case, is hardly guaranteed. To deal with such a problem, we propose a novel approach called Cooperative Training (CoT) to improve the training of sequence generative models. CoT transforms the min-max game of GANs into a joint maximization framework and manages to explicitly estimate and optimize Jensen-Shannon divergence. Moreover, CoT works without the necessity of pre-training via MLE, which is crucial to the success of previous methods. In the experiments, compared to existing state-of-the-art methods, CoT shows superior or at least competitive performance on sample quality, diversity, as well as training stability.
****
Generalized Approximate Survey Propagation for High-Dimensional Estimation
Carlo Lucibello, Luca Saglietti, Yue Lu
In Generalized Linear Estimation (GLE) problems, we seek to estimate a signal that is observed through a linear transform followed by a component-wise, possibly nonlinear and noisy, channel. In the Bayesian optimal setting, Generalized Approximate Message Passing (GAMP) is known to achieve optimal performance for GLE. However, its performance can significantly deteriorate whenever there is a mismatch between the assumed and the true generative model, a situation frequently encountered in practice. In this paper, we propose a new algorithm, named Generalized Approximate Survey Propagation (GASP), for solving GLE in the presence of prior or model misspecifications. As a prototypical example, we consider the phase retrieval problem, where we show that GASP outperforms the corresponding GAMP, reducing the reconstruction threshold and, for certain choices of its parameters, approaching Bayesian optimal performance. Furthermore, we present a set of state evolution equations that can precisely characterize the performance of GASP in the high-dimensional limit.
****
High-Fidelity Image Generation With Fewer Labels
Mario Lu■i■, Michael Tschannen, Marvin Ritter, Xiaohua Zhai, Olivier Bachem, Sylvain Gelly
Deep generative models are becoming a cornerstone of modern machine learning. Recent work on conditional generative adversarial networks has shown that learning complex, high-dimensional distributions over natural images is within reach. While the latest models are able to generate high-fidelity, diverse natural images at high resolution, they rely on a vast quantity of labeled data. In this work we demonstrate how one can benefit from recent work on self- and semi-supervised learning to outperform the state of the art on both unsupervised ImageNet synthesis, as well as in the conditional setting. In particular, the proposed approach is able to match the sample quality (as measured by FID) of the current state-of-the-art conditional model BigGAN on ImageNet using only 10% of the labels and outperform it using 20% of the labels.
****
Leveraging Low-Rank Relations Between Surrogate Tasks in Structured Prediction
Giulia Luise, Dimitrios Stamos, Massimiliano Pontil, Carlo Ciliberto
We study the interplay between surrogate methods for structured prediction and techniques from multitask learning designed to leverage relationships between surrogate outputs. We propose an efficient algorithm based on trace norm regularization which, differently from previous methods, does not require explicit knowledge of the coding/decoding functions of the surrogate framework. As a result, our algorithm can be applied to the broad class of problems in which the surrogate space is large or even infinite dimensional. We study excess risk bounds for trace norm regularized structured prediction proving the consistency and learning rates for our estimator. We also identify relevant regimes in which our approach

can enjoy better generalization performance than previous methods. Numerical experiments on ranking problems indicate that enforcing low-rank relations among surrogate outputs may indeed provide a significant advantage in practice.
****

Differentiable Dynamic Normalization for Learning Deep Representation
Ping Luo, Peng Zhanglin, Shao Wenqi, Zhang Ruimao, Ren Jiamin, Wu Lingyun
This work presents Dynamic Normalization (DN), which is able to learn arbitrary normalization operations for different convolutional layers in a deep ConvNet. Unlike existing normalization approaches that predefined computations of the statistics (mean and variance), DN learns to estimate them. DN has several appealing benefits. First, it adapts to various networks, tasks, and batch sizes. Second, it can be easily implemented and trained in a differentiable end-to-end manner with merely small number of parameters. Third, its matrix formulation represents a wide range of normalization methods, shedding light on analyzing them theoretically. Extensive studies show that DN outperforms its counterparts in CIFAR10 and ImageNet.
****

Disentangled Graph Convolutional Networks
Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, Wenwu Zhu
The formation of a real-world graph typically arises from the highly complex interaction of many latent factors. The existing deep learning methods for graph-structured data neglect the entanglement of the latent factors, rendering the learned representations non-robust and hardly explainable. However, learning representations that disentangle the latent factors poses great challenges and remains largely unexplored in the literature of graph neural networks. In this paper, we introduce the disentangled graph convolutional network (DisenGCN) to learn disentangled node representations. In particular, we propose a novel neighborhood routing mechanism, which is capable of dynamically identifying the latent factor that may have caused the edge between a node and one of its neighbors, and accordingly assigning the neighbor to a channel that extracts and convolutes features specific to that factor. We theoretically prove the convergence properties of the routing mechanism. Empirical results show that our proposed model can achieve significant performance gains, especially when the data demonstrate the existence of many entangled factors.
****

Variational Implicit Processes
Chao Ma, Yingzhen Li, Jose Miguel Hernandez-Lobato
We introduce the implicit processes (IPs), a stochastic process that places implicitly defined multivariate distributions over any finite collections of random variables. IPs are therefore highly flexible implicit priors over functions, with examples including data simulators, Bayesian neural networks and non-linear transformations of stochastic processes. A novel and efficient approximate inference algorithm for IPs, namely the variational implicit processes (VIPs), is derived using generalised wake-sleep updates. This method returns simple update equations and allows scalable hyper-parameter learning with stochastic optimization. Experiments show that VIPs return better uncertainty estimates and lower errors over existing inference methods for challenging models such as Bayesian neural networks, and Gaussian processes.
****

EDDI: Efficient Dynamic Discovery of High-Value Information with Partial VAE
Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez-Lobato, Sebastian Nowozin, Cheng Zhang
Many real-life decision making situations allow further relevant information to be acquired at a specific cost, for example, in assessing the health status of a patient we may decide to take additional measurements such as diagnostic tests or imaging scans before making a final assessment. Acquiring more relevant information enables better decision making, but may be costly. How can we trade off the desire to make good decisions by acquiring further information with the cost of performing that acquisition? To this end, we propose a principled framework, named EDDI (Efficient Dynamic Discovery of high-value Information), based on the

theory of Bayesian experimental design. In EDDI, we propose a novel partial variational autoencoder (Partial VAE) to predict missing data entries problematically given any subset of the observed ones, and combine it with an acquisition function that maximizes expected information gain on a set of target variables. We show cost reduction at the same decision quality and improved decision quality at the same cost in multiple machine learning benchmarks and two real-world health-care applications.
****
Bayesian leave-one-out cross-validation for large data
Måns Magnusson, Michael Andersen, Johan Jonasson, Aki Vehtari
Model inference, such as model comparison, model checking, and model selection, is an important part of model development. Leave-one-out cross-validation (LOO) is a general approach for assessing the generalizability of a model, but unfortunately, LOO does not scale well to large datasets. We propose a combination of using approximate inference techniques and probability-proportional-to-size-sampling (PPS) for fast LOO model evaluation for large datasets. We provide both theoretical and empirical results showing good properties for large data.
****
Composable Core-sets for Determinant Maximization: A Simple Near-Optimal Algorithm
Sepideh Mahabadi, Piotr Indyk, Shayan Oveis Gharan, Alireza Rezaei
“Composable core-sets” are an efficient framework for solving optimization problems in massive data models. In this work, we consider efficient construction of composable core-sets for the determinant maximization problem. This can also be cast as the MAP inference task for “determinantal point processes", that have recently gained a lot of interest for modeling diversity and fairness. The problem was recently studied in \cite{indyk2018composable}, where they designed composable core-sets with the optimal approximation bound of $O(k)^k$. On the other hand, the more practical “Greedy" algorithm has been previously used in similar contexts. In this work, first we provide a theoretical approximation guarantee of C^{k^2} for the Greedy algorithm in the context of composable core-sets; Further, we propose to use a “Local Search" based algorithm that while being still practical, achieves a nearly optimal approximation bound of $O(k)^{2k}$; Finally, we implement all three algorithms and show the effectiveness of our proposed algorithm on standard data sets.
****
Guided evolutionary strategies: augmenting random search with surrogate gradients
Niru Maheswaranathan, Luke Metz, George Tucker, Dami Choi, Jascha Sohl-Dickstein
Many applications in machine learning require optimizing a function whose true gradient is unknown or computationally expensive, but where surrogate gradient information, directions that may be correlated with the true gradient, is cheaply available. For example, this occurs when an approximate gradient is easier to compute than the full gradient (e.g. in meta-learning or unrolled optimization), or when a true gradient is intractable and is replaced with a surrogate (e.g. in reinforcement learning or training networks with discrete variables). We propose Guided Evolutionary Strategies (GES), a method for optimally using surrogate gradient directions to accelerate random search. GES defines a search distribution for evolutionary strategies that is elongated along a subspace spanned by the surrogate gradients and estimates a descent direction which can then be passed to a first-order optimizer. We analytically and numerically characterize the tradeoffs that result from tuning how strongly the search distribution is stretched along the guiding subspace and use this to derive a setting of the hyperparameters that works well across problems. We evaluate GES on several example problems, demonstrating an improvement over both standard evolutionary strategies and first-order methods that directly follow the surrogate gradient.
****
Universal Multi-Party Poisoning Attacks
Saeed Mahloujifar, Mohammad Mahmoody, Ameer Mohammed
In this work, we demonstrate universal multi-party poisoning attacks that adapt

and apply to any multi-party learning process with arbitrary interaction pattern
 between the parties. More generally, we introduce and study (k,p)-poisoning a
ttacks in which an adversary controls $k\in[m]$ of the parties, and for each cor
rupted party P_i, the adversary submits some poisoned data T'_i on behalf of
 P_i that is still "$(1-p)$-close" to the correct data T_i (e.g., $1-p$ frac
tion of T'_i is still honestly generated).We prove that for any "bad" property
 B of the final trained hypothesis h (e.g., h failing on a particular test
 example or having "large" risk) that has an arbitrarily small constant probabil
ity of happening without the attack, there always is a (k,p)-poisoning attack
that increases the probability of B from μ to by $\mu^{1-p \cdot k/m} = \m
u + \Omega(p \cdot k/m)$. Our attack only uses clean labels, and it is online, a
s it only knows the the data shared so far.
****
Traditional and Heavy Tailed Self Regularization in Neural Network Models
Michael Mahoney, Charles Martin
Random Matrix Theory (RMT) is applied to analyze the weight matrices of Deep Neu
ral Networks (DNNs), including both production quality, pre-trained models such
as AlexNet and Inception, and smaller models trained from scratch, such as LeNet
5 and a miniature-AlexNet. Empirical and theoretical results clearly indicate th
at the empirical spectral density (ESD) of DNN layer matrices displays signature
s of traditionally-regularized statistical models, even in the absence of exogen
ously specifying traditional forms of regularization, such as Dropout or Weight
Norm constraints. Building on recent results in RMT, most notably its extension
to Universality classes of Heavy-Tailed matrices, we develop a theory to identif
y 5+1 Phases of Training, corresponding to increasing amounts of Implicit Self-R
egularization. For smaller and/or older DNNs, this Implicit Self-Regularization
is like traditional Tikhonov regularization, in that there is a “size scale” sep
arating signal from noise. For state-of-the-art DNNs, however, we identify a nov
el form of Heavy-Tailed Self-Regularization, similar to the self-organization se
en in the statistical physics of disordered systems. This implicit Self-Regulari
zation can depend strongly on the many knobs of the training process. By exploit
ing the generalization gap phenomena, we demonstrate that we can cause a small m
odel to exhibit all 5+1 phases of training simply by changing the batch size.
****
Curvature-Exploiting Acceleration of Elastic Net Computations
Vien Mai, Mikael Johansson
This paper introduces an efficient second-order method for solving the elastic n
et problem. Its key innovation is a computationally efficient technique for inje
cting curvature information in the optimization process which admits a strong th
eoretical performance guarantee. In particular, we show improved run time over p
opular first-order methods and quantify the speed-up in terms of statistical mea
sures of the data matrix. The improved time complexity is the result of an exten
sive exploitation of the problem structure and a careful combination of second-o
rder information, variance reduction techniques, and momentum acceleration. Besi
de theoretical speed-up, experimental results demonstrate great practical perfor
mance benefits of curvature information, especially for ill-conditioned data set
s.
****
Breaking the gridlock in Mixture-of-Experts: Consistent and Efficient Algorithms
Ashok Makkuva, Pramod Viswanath, Sreeram Kannan, Sewoong Oh
Mixture-of-Experts (MoE) is a widely popular model for ensemble learning and is
a basic building block of highly successful modern neural networks as well as a
component in Gated Recurrent Units (GRU) and Attention networks. However, presen
t algorithms for learning MoE, including the EM algorithm and gradient descent,
are known to get stuck in local optima. From a theoretical viewpoint, finding an
 efficient and provably consistent algorithm to learn the parameters remains a l
ong standing open problem for more than two decades. In this paper, we introduce
 the first algorithm that learns the true parameters of a MoE model for a wide c
lass of non-linearities with global consistency guarantees. While existing algor
ithms jointly or iteratively estimate the expert parameters and the gating param

eters in the MoE, we propose a novel algorithm that breaks the deadlock and can directly estimate the expert parameters by sensing its echo in a carefully designed cross-moment tensor between the inputs and the output. Once the experts are known, the recovery of gating parameters still requires an EM algorithm; however, we show that the EM algorithm for this simplified problem, unlike the joint EM algorithm, converges to the true parameters. We empirically validate our algorithm on both the synthetic and real data sets in a variety of settings, and show superior performance to standard baselines.
****

Calibrated Model-Based Deep Reinforcement Learning
Ali Malik, Volodymyr Kuleshov, Jiaming Song, Danny Nemer, Harlan Seymour, Stefano Ermon
Estimates of predictive uncertainty are important for accurate model-based planning and reinforcement learning. However, predictive uncertainties — especially ones derived from modern deep learning systems — can be inaccurate and impose a bottleneck on performance. This paper explores which uncertainties are needed for model-based reinforcement learning and argues that ideal uncertainties should be calibrated, i.e. their probabilities should match empirical frequencies of predicted events. We describe a simple way to augment any model-based reinforcement learning agent with a calibrated model and show that doing so consistently improves planning, sample complexity, and exploration. On the \textsc{HalfCheetah} MuJoCo task, our system achieves state-of-the-art performance using 50% fewer samples than the current leading approach. Our findings suggest that calibration can improve the performance of model-based reinforcement learning with minimal computational and implementation overhead.
****

Learning from Delayed Outcomes via Proxies with Applications to Recommender Systems
Timothy Arthur Mann, Sven Gowal, Andras Gyorgy, Huiyi Hu, Ray Jiang, Balaji Lakshminarayanan, Prav Srinivasan
Predicting delayed outcomes is an important problem in recommender systems (e.g., if customers will finish reading an ebook). We formalize the problem as an adversarial, delayed online learning problem and consider how a proxy for the delayed outcome (e.g., if customers read a third of the book in 24 hours) can help minimize regret, even though the proxy is not available when making a prediction. Motivated by our regret analysis, we propose two neural network architectures: Factored Forecaster (FF) which is ideal if the proxy is informative of the outcome in hindsight, and Residual Factored Forecaster (RFF) that is robust to a non-informative proxy. Experiments on two real-world datasets for predicting human behavior show that RFF outperforms both FF and a direct forecaster that does not make use of the proxy. Our results suggest that exploiting proxies by factorization is a promising way to mitigate the impact of long delays in human-behavior prediction tasks.
****

Passed & Spurious: Descent Algorithms and Local Minima in Spiked Matrix-Tensor Models
Stefano Sarao Mannelli, Florent Krzakala, Pierfrancesco Urbani, Lenka Zdeborova
In this work we analyse quantitatively the interplay between the loss landscape and performance of descent algorithms in a prototypical inference problem, the spiked matrix-tensor model. We study a loss function that is the negative log-likelihood of the model. We analyse the number of local minima at a fixed distance from the signal/spike with the Kac-Rice formula, and locate trivialization of the landscape at large signal-to-noise ratios. We evaluate analytically the performance of a gradient flow algorithm using integro-differential PDEs as developed in physics of disordered systems for the Langevin dynamics. We analyze the performance of an approximate message passing algorithm estimating the maximum likelihood configuration via its state evolution. We conclude by comparing the above results: while we observe a drastic slow down of the gradient flow dynamics even in the region where the landscape is trivial, both the analyzed algorithms are shown to perform well even in the part of the region of parameters where spurious

local minima are present.
****
A Baseline for Any Order Gradient Estimation in Stochastic Computation Graphs
Jingkai Mao, Jakob Foerster, Tim Rocktäschel, Maruan Al-Shedivat, Gregory Farquhar, Shimon Whiteson
By enabling correct differentiation in Stochastic Computation Graphs (SCGs), the infinitely differentiable Monte-Carlo estimator (DiCE) can generate correct estimates for the higher order gradients that arise in, e.g., multi-agent reinforcement learning and meta-learning. However, the baseline term in DiCE that serves as a control variate for reducing variance applies only to first order gradient estimation, limiting the utility of higher-order gradient estimates. To improve the sample efficiency of DiCE, we propose a new baseline term for higher order gradient estimation. This term may be easily included in the objective, and produces unbiased variance-reduced estimators under (automatic) differentiation, without affecting the estimate of the objective itself or of the first order gradient estimate. It reuses the same baseline function (e.g., the state-value function in reinforcement learning) already used for the first order baseline. We provide theoretical analysis and numerical evaluations of this new baseline, which demonstrate that it can dramatically reduce the variance of DiCE’s second order gradient estimators and also show empirically that it reduces the variance of third and fourth order gradients. This computational tool can be easily used to estimate higher order gradients with unprecedented efficiency and simplicity wherever automatic differentiation is utilised, and it has the potential to unlock applications of higher order gradients in reinforcement learning and meta-learning.
****
Adversarial Generation of Time-Frequency Features with application in audio synthesis
Andrés Marafioti, Nathanaël Perraudin, Nicki Holighaus, Piotr Majdak
Time-frequency (TF) representations provide powerful and intuitive features for the analysis of time series such as audio. But still, generative modeling of audio in the TF domain is a subtle matter. Consequently, neural audio synthesis widely relies on directly modeling the waveform and previous attempts at unconditionally synthesizing audio from neurally generated invertible TF features still struggle to produce audio at satisfying quality. In this article, focusing on the short-time Fourier transform, we discuss the challenges that arise in audio synthesis based on generated invertible TF features and how to overcome them. We demonstrate the potential of deliberate generative TF modeling by training a generative adversarial network (GAN) on short-time Fourier features. We show that by applying our guidelines, our TF-based network was able to outperform a state-of-the-art GAN generating waveforms directly, despite the similar architecture in the two networks.
****
On the Universality of Invariant Networks
Haggai Maron, Ethan Fetaya, Nimrod Segol, Yaron Lipman
Constraining linear layers in neural networks to respect symmetry transformations from a group G is a common design principle for invariant networks that has found many applications in machine learning. ■■ In this paper, we consider a fundamental question that has received very little attention to date: Can these networks approximate any (continuous) invariant function? ■■ We tackle the rather general case where $G\leq S_n$ (an arbitrary subgroup of the symmetric group) that acts on $\mathbb{R}^n$ by permuting coordinates. This setting includes several recent popular invariant networks. We present two main results: First, G-invariant networks are universal if high-order tensors are allowed. Second, there are groups G for which higher-order tensors are unavoidable for obtaining universality. ■■ G-invariant networks consisting of only first-order tensors are of special interest due to their practical value. We conclude the paper by proving a necessary condition for the universality of G-invariant networks that incorporate only first-order tensors. Lastly, we propose a conjecture stating that this condition is also sufficient.
****

Decomposing feature-level variation with Covariate Gaussian Process Latent Variable Models

Kaspar Märtens, Kieran Campbell, Christopher Yau

The interpretation of complex high-dimensional data typically requires the use of dimensionality reduction techniques to extract explanatory low-dimensional representations. However, in many real-world problems these representations may not be sufficient to aid interpretation on their own, and it would be desirable to interpret the model in terms of the original features themselves. Our goal is to characterise how feature-level variation depends on latent low-dimensional representations, external covariates, and non-linear interactions between the two. In this paper, we propose to achieve this through a structured kernel decomposition in a hybrid Gaussian Process model which we call the Covariate Gaussian Process Latent Variable Model (c-GPLVM). We demonstrate the utility of our model on simulated examples and applications in disease progression modelling from high-dimensional gene expression data in the presence of additional phenotypes. In each setting we show how the c-GPLVM can extract low-dimensional structures from high-dimensional data sets whilst allowing a breakdown of feature-level variability that is not present in other commonly used dimensionality reduction approaches.

****

Fairness-Aware Learning for Continuous Attributes and Treatments

Jeremie Mary, Clément Calauzènes, Noureddine El Karoui

We address the problem of algorithmic fairness: ensuring that the outcome of a classifier is not biased towards certain values of sensitive variables such as age, race or gender. As common fairness metrics can be expressed as measures of (conditional) independence between variables, we propose to use the Rényi maximum correlation coefficient to generalize fairness measurement to continuous variables. We exploit Witsenhausen’s characterization of the Rényi correlation coefficient to propose a differentiable implementation linked to f-divergences. This allows us to generalize fairness-aware learning to continuous variables by using a penalty that upper bounds this coefficient. Theses allows fairness to be extented to variables such as mixed ethnic groups or financial status without thresholds effects. This penalty can be estimated on mini-batches allowing to use deep nets. Experiments show favorable comparisons to state of the art on binary variables and prove the ability to protect continuous ones

****

Optimal Minimal Margin Maximization with Boosting

Alexander Mathiasen, Kasper Green Larsen, Allan Grønlund

Boosting algorithms iteratively produce linear combinations of more and more base hypotheses and it has been observed experimentally that the generalization error keeps improving even after achieving zero training error. One popular explanation attributes this to improvements in margins. A common goal in a long line of research, is to obtain large margins using as few base hypotheses as possible, culminating with the AdaBoostV algorithm by R{ä}tsch and Warmuth [JMLR’05]. The AdaBoostV algorithm was later conjectured to yield an optimal trade-off between number of hypotheses trained and the minimal margin over all training points (Nie, Warmuth, Vishwanathan and Zhang [JMLR’13]). Our main contribution is a new algorithm refuting this conjecture. Furthermore, we prove a lower bound which implies that our new algorithm is optimal.

****

Disentangling Disentanglement in Variational Autoencoders

Emile Mathieu, Tom Rainforth, N Siddharth, Yee Whye Teh

We develop a generalisation of disentanglement in variational autoencoders (VAEs)—decomposition of the latent representation—characterising it as the fulfilment of two factors: a) the latent encodings of the data having an appropriate level of overlap, and b) the aggregate encoding of the data conforming to a desired structure, represented through the prior. Decomposition permits disentanglement, i.e. explicit independence between latents, as a special case, but also allows for a much richer class of properties to be imposed on the learnt representation, such as sparsity, clustering, independent subspaces, or even intricate hierarchical dependency relationships. We show that the β-VAE varies from the stan

dard VAE predominantly in its control of latent overlap and that for the standar
d choice of an isotropic Gaussian prior, its objective is invariant to rotations
 of the latent representation. Viewed from the decomposition perspective, breaki
ng this invariance with simple manipulations of the prior can yield better disen
tanglement with little or no detriment to reconstructions. We further demonstrat
e how other choices of prior can assist in producing different decompositions an
d introduce an alternative training objective that allows the control of both de
composition factors in a principled manner.
****
MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets
Pierre-Alexandre Mattei, Jes Frellsen
We consider the problem of handling missing data with deep latent variable model
s (DLVMs). First, we present a simple technique to train DLVMs when the training
 set contains missing-at-random data. Our approach, called MIWAE, is based on th
e importance-weighted autoencoder (IWAE), and maximises a potentially tight lowe
r bound of the log-likelihood of the observed data. Compared to the original IWA
E, our algorithm does not induce any additional computational overhead due to th
e missing data. We also develop Monte Carlo techniques for single and multiple i
mputation using a DLVM trained on an incomplete data set. We illustrate our appr
oach by training a convolutional DLVM on incomplete static binarisations of MNIS
T. Moreover, on various continuous data sets, we show that MIWAE provides extrem
ely accurate single imputations, and is highly competitive with state-of-the-art
 methods.
****
Distributional Reinforcement Learning for Efficient Exploration
Borislav Mavrin, Hengshuai Yao, Linglong Kong, Kaiwen Wu, Yaoliang Yu
In distributional reinforcement learning (RL), the estimated distribution of val
ue functions model both the parametric and intrinsic uncertainties. We propose a
 novel and efficient exploration method for deep RL that has two components. The
 first is a decaying schedule to suppress the intrinsic uncertainty. The second
is an exploration bonus calculated from the upper quantiles of the learned distr
ibution. In Atari 2600 games, our method achieves 483 % average gain across 49 g
ames in cumulative rewards over QR-DQN. We also compared our algorithm with QR-D
QN in a challenging 3D driving simulator (CARLA). Results show that our algorith
m achieves nearoptimal safety rewards twice faster than QRDQN.
****
Graphical-model based estimation and inference for differential privacy
Ryan Mckenna, Daniel Sheldon, Gerome Miklau
Many privacy mechanisms reveal high-level information about a data distribution
through noisy measurements. It is common to use this information to estimate the
 answers to new queries. In this work, we provide an approach to solve this esti
mation problem efficiently using graphical models, which is particularly effecti
ve when the distribution is high-dimensional but the measurements are over low-d
imensional marginals. We show that our approach is far more efficient than exist
ing estimation techniques from the privacy literature and that it can improve th
e accuracy and scalability of many state-of-the-art mechanisms.
****
Efficient Amortised Bayesian Inference for Hierarchical and Nonlinear Dynamical
Systems
Geoffrey Roeder, Paul Grant, Andrew Phillips, Neil Dalchau, Edward Meeds
We introduce a flexible, scalable Bayesian inference framework for nonlinear dyn
amical systems characterised by distinct and hierarchical variability at the ind
ividual, group, and population levels. Our model class is a generalisation of no
nlinear mixed-effects (NLME) dynamical systems, the statistical workhorse for ma
ny experimental sciences. We cast parameter inference as stochastic optimisation
 of an end-to-end differentiable, block-conditional variational autoencoder. We
specify the dynamics of the data-generating process as an ordinary differential
equation (ODE) such that both the ODE and its solver are fully differentiable. T
his model class is highly flexible: the ODE right-hand sides can be a mixture of
 user-prescribed or "white-box" sub-components and neural network or "black-box"

sub-components. Using stochastic optimisation, our amortised inference algorithm could seamlessly scale up to massive data collection pipelines (common in labs with robotic automation). Finally, our framework supports interpretability with respect to the underlying dynamics, as well as predictive generalization to unseen combinations of group components (also called “zero-shot" learning). We empirically validate our method by predicting the dynamic behaviour of bacteria that were genetically engineered to function as biosensors.
****
Toward Controlling Discrimination in Online Ad Auctions
Elisa Celis, Anay Mehrotra, Nisheeth Vishnoi
Online advertising platforms are thriving due to the customizable audiences they offer advertisers. However, recent studies show that advertisements can be discriminatory with respect to the gender or race of the audience that sees the ad, and may inadvertently cross ethical and/or legal boundaries. To prevent this, we propose a constrained ad auction framework that maximizes the platform’s revenue conditioned on ensuring that the audience seeing an advertiser’s ad is distributed appropriately across sensitive types such as gender or race. Building upon Myerson’s classic work, we first present an optimal auction mechanism for a large class of fairness constraints. Finding the parameters of this optimal auction, however, turns out to be a non-convex problem. We show that this non-convex problem can be reformulated as a more structured non-convex problem with no saddle points or local-maxima; this allows us to develop a gradient-descent-based algorithm to solve it. Our empirical results on the A1 Yahoo! dataset demonstrate that our algorithm can obtain uniform coverage across different user types for each advertiser at a minor loss to the revenue of the platform, and a small change to the size of the audience each advertiser reaches.
****
Stochastic Blockmodels meet Graph Neural Networks
Nikhil Mehta, Lawrence Carin Duke, Piyush Rai
Stochastic blockmodels (SBM) and their variants, $e.g.$, mixed-membership and overlapping stochastic blockmodels, are latent variable based generative models for graphs. They have proven to be successful for various tasks, such as discovering the community structure and link prediction on graph-structured data. Recently, graph neural networks, $e.g.$, graph convolutional networks, have also emerged as a promising approach to learn powerful representations (embeddings) for the nodes in the graph, by exploiting graph properties such as locality and invariance. In this work, we unify these two directions by developing a sparse variational autoencoder for graphs, that retains the interpretability of SBMs, while also enjoying the excellent predictive performance of graph neural nets. Moreover, our framework is accompanied by a fast recognition model that enables fast inference of the node embeddings (which are of independent interest for inference in SBM and its variants). Although we develop this framework for a particular type of SBM, namely the overlapping stochastic blockmodel, the proposed framework can be adapted readily for other types of SBMs. Experimental results on several benchmarks demonstrate encouraging results on link prediction while learning an interpretable latent structure that can be used for community discovery.
****
Imputing Missing Events in Continuous-Time Event Streams
Hongyuan Mei, Guanghui Qin, Jason Eisner
Events in the world may be caused by other, unobserved events. We consider sequences of events in continuous time. Given a probability model of complete sequences, we propose particle smoothing—a form of sequential importance sampling—to impute the missing events in an incomplete sequence. We develop a trainable family of proposal distributions based on a type of bidirectional continuous-time LSTM: Bidirectionality lets the proposals condition on future observations, not just on the past as in particle filtering. Our method can sample an ensemble of possible complete sequences (particles), from which we form a single consensus prediction that has low Bayes risk under our chosen loss metric. We experiment in multiple synthetic and real domains, using different missingness mechanisms, and modeling the complete sequences in each domain with a neural Hawkes process (Mei &

Eisner 2017). On held-out incomplete sequences, our method is effective at infe
rring the ground-truth unobserved events, with particle smoothing consistently i
mproving upon particle filtering.
****
Same, Same But Different: Recovering Neural Network Quantization Error Through W
eight Factorization
Eldad Meller, Alexander Finkelstein, Uri Almog, Mark Grobman
Quantization of neural networks has become common practice, driven by the need f
or efficient implementations of deep neural networks on embedded devices. In thi
s paper, we exploit an oft-overlooked degree of freedom in most networks - for a
 given layer, individual output channels can be scaled by any factor provided th
at the corresponding weights of the next layer are inversely scaled. Therefore,
a given network has many factorizations which change the weights of the network
without changing its function. We present a conceptually simple and easy to impl
ement method that uses this property and show that proper factorizations signifi
cantly decrease the degradation caused by quantization. We show improvement on a
 wide variety of networks and achieve state-of-the-art degradation results for M
obileNets. While our focus is on quantization, this type of factorization is app
licable to other domains such as network-pruning, neural nets regularization and
 network interpretability.
****
The Wasserstein Transform
Facundo Memoli, Zane Smith, Zhengchao Wan
We introduce the Wasserstein transform, a method for enhancing and denoising dat
asets defined on general metric spaces. The construction draws inspiration from
Optimal Transportation ideas. We establish the stability of our method under dat
a perturbation and, when the dataset is assumed to be Euclidean, we also exhibit
 a precise connection between the Wasserstein transform and the mean shift famil
y of algorithms. We then use this connection to prove that mean shift also inher
its stability under perturbations. We study the performance of the Wasserstein t
ransform method on different datasets as a preprocessing step prior to clusterin
g and classification tasks.
****
Ithemal: Accurate, Portable and Fast Basic Block Throughput Estimation using Dee
p Neural Networks
Charith Mendis, Alex Renda, Dr.Saman Amarasinghe, Michael Carbin
Predicting the number of clock cycles a processor takes to execute a block of as
sembly instructions in steady state (the throughput) is important for both compi
ler designers and performance engineers. Building an analytical model to do so i
s especially complicated in modern x86-64 Complex Instruction Set Computer (CISC
) machines with sophisticated processor microarchitectures in that it is tedious
, error prone, and must be performed from scratch for each processor generation.
 In this paper we present Ithemal, the first tool which learns to predict the th
roughput of a set of instructions. Ithemal uses a hierarchical LSTM–based approa
ch to predict throughput based on the opcodes and operands of instructions in a
basic block. We show that Ithemal is more accurate than state-of-the-art hand-wr
itten tools currently used in compiler backends and static machine code analyzer
s. In particular, our model has less than half the error of state-of-the-art ana
lytical models (LLVM’s llvm-mca and Intel’s IACA). Ithemal is also able to predi
ct these throughput values just as fast as the aforementioned tools, and is easi
ly ported across a variety of processor microarchitectures with minimal develope
r effort.
****
Geometric Losses for Distributional Learning
Arthur Mensch, Mathieu Blondel, Gabriel Peyré
Building upon recent advances in entropy-regularized optimal transport, and upon
 Fenchel duality between measures and continuous functions, we propose a general
ization of the logistic loss that incorporates a metric or cost between classes.
 Unlike previous attempts to use optimal transport distances for learning, our l
oss results in unconstrained convex objective functions, supports infinite (or v

ery large) class spaces, and naturally defines a geometric generalization of the softmax operator. The geometric properties of this loss make it suitable for predicting sparse and singular distributions, for instance supported on curves or hyper-surfaces. We study the theoretical properties of our loss and showcase its effectiveness on two applications: ordinal regression and drawing generation.
****
Spectral Clustering of Signed Graphs via Matrix Power Means
Pedro Mercado, Francesco Tudisco, Matthias Hein
Signed graphs encode positive (attractive) and negative (repulsive) relations between nodes. We extend spectral clustering to signed graphs via the one-parameter family of Signed Power Mean Laplacians, defined as the matrix power mean of normalized standard and signless Laplacians of positive and negative edges. We provide a thorough analysis of the proposed approach in the setting of a general Stochastic Block Model that includes models such as the Labeled Stochastic Block Model and the Censored Block Model. We show that in expectation the signed power mean Laplacian captures the ground truth clusters under reasonable settings where state-of-the-art approaches fail. Moreover, we prove that the eigenvalues and eigenvector of the signed power mean Laplacian concentrate around their expectation under reasonable conditions in the general Stochastic Block Model. Extensive experiments on random graphs and real world datasets confirm the theoretically predicted behaviour of the signed power mean Laplacian and show that it compares favourably with state-of-the-art methods.
****
Simple Stochastic Gradient Methods for Non-Smooth Non-Convex Regularized Optimization
Michael Metel, Akiko Takeda
Our work focuses on stochastic gradient methods for optimizing a smooth non-convex loss function with a non-smooth non-convex regularizer. Research on this class of problem is quite limited, and until recently no non-asymptotic convergence results have been reported. We present two simple stochastic gradient algorithms, for finite-sum and general stochastic optimization problems, which have superior convergence complexities compared to the current state-of-the-art. We also compare our algorithms’ performance in practice for empirical risk minimization.
****
Reinforcement Learning in Configurable Continuous Environments
Alberto Maria Metelli, Emanuele Ghelfi, Marcello Restelli
Configurable Markov Decision Processes (Conf-MDPs) have been recently introduced as an extension of the usual MDP model to account for the possibility of configuring the environment to improve the agent’s performance. Currently, there is still no suitable algorithm to solve the learning problem for real-world Conf-MDPs. In this paper, we fill this gap by proposing a trust-region method, Relative Entropy Model Policy Search (REMPS), able to learn both the policy and the MDP configuration in continuous domains without requiring the knowledge of the true model of the environment. After introducing our approach and providing a finite-sample analysis, we empirically evaluate REMPS on both benchmark and realistic environments by comparing our results with those of the gradient methods.
****
Understanding and correcting pathologies in the training of learned optimizers
Luke Metz, Niru Maheswaranathan, Jeremy Nixon, Daniel Freeman, Jascha Sohl-Dickstein
Deep learning has shown that learned functions can dramatically outperform hand-designed functions on perceptual tasks. Analogously, this suggests that learned optimizers may similarly outperform current hand-designed optimizers, especially for specific problems. However, learned optimizers are notoriously difficult to train and have yet to demonstrate wall-clock speedups over hand-designed optimizers, and thus are rarely used in practice. Typically, learned optimizers are trained by truncated backpropagation through an unrolled optimization process. The resulting gradients are either strongly biased (for short truncations) or have exploding norm (for long truncations). In this work we propose a training scheme which overcomes both of these difficulties, by dynamically weighting two unbias

ed gradient estimators for a variational loss on optimizer performance. This all
ows us to train neural networks to perform optimization of a specific task faste
r than tuned first-order methods. Moreover, by training the optimizer against va
lidation loss (as opposed to training loss), we are able to learn optimizers tha
t train networks to generalize better than first order methods. We demonstrate t
hese results on problems where our learned optimizer trains convolutional networ
ks faster in wall-clock time compared to tuned first-order methods and with an i
mprovement in test loss.
****
Optimality Implies Kernel Sum Classifiers are Statistically Efficient
Raphael Meyer, Jean Honorio
We propose a novel combination of optimization tools with learning theory bounds
 in order to analyze the sample complexity of optimal kernel sum classifiers. Th
is contrasts the typical learning theoretic results which hold for all (potentia
lly suboptimal) classifiers. Our work also justifies assumptions made in prior w
ork on multiple kernel learning. As a byproduct of our analysis, we also provide
 a new form of Rademacher complexity for hypothesis classes containing only opti
mal classifiers.
****
On Dropout and Nuclear Norm Regularization
Poorya Mianjy, Raman Arora
We give a formal and complete characterization of the explicit regularizer induc
ed by dropout in deep linear networks with squared loss. We show that (a) the ex
plicit regularizer is composed of an ℓ_2-path regularizer and other terms t
hat are also re-scaling invariant, (b) the convex envelope of the induced regula
rizer is the squared nuclear norm of the network map, and (c) for a sufficiently
 large dropout rate, we characterize the global optima of the dropout objective.
 We validate our theoretical findings with empirical results.
****
Discriminative Regularization for Latent Variable Models with Applications to El
ectrocardiography
Andrew Miller, Ziad Obermeyer, John Cunningham, Sendhil Mullainathan
Generative models often use latent variables to represent structured variation i
n high-dimensional data, such as images and medical waveforms. However, these la
tent variables may ignore subtle, yet meaningful features in the data. Some feat
ures may predict an outcome of interest (e.g. heart attack) but account for only
 a small fraction of variation in the data. We propose a generative model traini
ng objective that uses a black-box discriminative model as a regularizer to lear
n representations that preserve this predictive variation. With these discrimina
tively regularized latent variable models, we visualize and measure variation in
 the data that influence a black-box predictive model, enabling an expert to bet
ter understand each prediction. With this technique, we study models that use el
ectrocardiograms to predict outcomes of clinical interest. We measure our approa
ch on synthetic and real data with statistical summaries and an experiment carri
ed out by a physician.
****
Formal Privacy for Functional Data with Gaussian Perturbations
Ardalan Mirshani, Matthew Reimherr, Aleksandra Slavkovi■
Motivated by the rapid rise in statistical tools in Functional Data Analysis, we
 consider the Gaussian mechanism for achieving differential privacy (DP) with pa
rameter estimates taking values in a, potentially infinite-dimensional, separabl
e Banach space. Using classic results from probability theory, we show how densi
ties over function spaces can be utilized to achieve the desired DP bounds. This
 extends prior results of Hall et al (2013) to a much broader class of statistic
al estimates and summaries, including “path level" summaries, nonlinear function
als, and full function releases. By focusing on Banach spaces, we provide a deep
er picture of the challenges for privacy with complex data, especially the role
regularization plays in balancing utility and privacy. Using an application to p
enalized smoothing, we highlight this balance in the context of mean function es
timation. Simulations and an application to {diffusion tensor imaging} are brief

ly presented, with extensive additions included in a supplement.
****
Co-manifold learning with missing data
Gal Mishne, Eric Chi, Ronald Coifman
Representation learning is typically applied to only one mode of a data matrix, either its rows or columns. Yet in many applications, there is an underlying geometry to both the rows and the columns. We propose utilizing this coupled structure to perform co-manifold learning: uncovering the underlying geometry of both the rows and the columns of a given matrix, where we focus on a missing data setting. Our unsupervised approach consists of three components. We first solve a family of optimization problems to estimate a complete matrix at multiple scales of smoothness. We then use this collection of smooth matrix estimates to compute pairwise distances on the rows and columns based on a new multi-scale metric that implicitly introduces a coupling between the rows and the columns. Finally, we construct row and column representations from these multi-scale metrics. We demonstrate that our approach outperforms competing methods in both data visualization and clustering.
****
Agnostic Federated Learning
Mehryar Mohri, Gary Sivek, Ananda Theertha Suresh
A key learning scenario in large-scale applications is that of federated learning, where a centralized model is trained based on data originating from a large number of clients. We argue that, with the existing training and inference, federated models can be biased towards different clients. Instead, we propose a new framework of agnostic federated learning, where the centralized model is optimized for any target distribution formed by a mixture of the client distributions. We further show that this framework naturally yields a notion of fairness. We present data-dependent Rademacher complexity guarantees for learning with this objective, which guide the definition of an algorithm for agnostic federated learning. We also give a fast stochastic optimization algorithm for solving the corresponding optimization problem, for which we prove convergence bounds, assuming a convex loss function and a convex hypothesis set. We further empirically demonstrate the benefits of our approach in several datasets. Beyond federated learning, our framework and algorithm can be of interest to other learning scenarios such as cloud computing, domain adaptation, drifting, and other contexts where the training and test distributions do not coincide.
****
Flat Metric Minimization with Applications in Generative Modeling
Thomas Möllenhoff, Daniel Cremers
We take the novel perspective to view data not as a probability distribution but rather as a current. Primarily studied in the field of geometric measure theory, k-currents are continuous linear functionals acting on compactly supported smooth differential forms and can be understood as a generalized notion of oriented k-dimensional manifold. By moving from distributions (which are 0-currents) to k-currents, we can explicitly orient the data by attaching a k-dimensional tangent plane to each sample point. Based on the flat metric which is a fundamental distance between currents, we derive FlatGAN, a formulation in the spirit of generative adversarial networks but generalized to k-currents. In our theoretical contribution we prove that the flat metric between a parametrized current and a reference current is Lipschitz continuous in the parameters. In experiments, we show that the proposed shift to k>0 leads to interpretable and disentangled latent representations which behave equivariantly to the specified oriented tangent planes.
****
Parsimonious Black-Box Adversarial Attacks via Efficient Combinatorial Optimization
Seungyong Moon, Gaon An, Hyun Oh Song
Solving for adversarial examples with projected gradient descent has been demonstrated to be highly effective in fooling the neural network based classifiers. However, in the black-box setting, the attacker is limited only to the query acce

ss to the network and solving for a successful adversarial example becomes much more difficult. To this end, recent methods aim at estimating the true gradient signal based on the input queries but at the cost of excessive queries. We propose an efficient discrete surrogate to the optimization problem which does not require estimating the gradient and consequently becomes free of the first order update hyperparameters to tune. Our experiments on Cifar-10 and ImageNet show the state of the art black-box attack performance with significant reduction in the required queries compared to a number of recently proposed methods. The source code is available at https://github.com/snu-mllab/parsimonious-blackbox-attack.

****

Parameter efficient training of deep convolutional neural networks by dynamic sparse reparameterization

Hesham Mostafa, Xin Wang

Modern deep neural networks are typically highly overparameterized. Pruning techniques are able to remove a significant fraction of network parameters with little loss in accuracy. Recently, techniques based on dynamic reallocation of non-zero parameters have emerged, allowing direct training of sparse networks without having to pre-train a large dense model. Here we present a novel dynamic sparse reparameterization method that addresses the limitations of previous techniques such as high computational cost and the need for manual configuration of the number of free parameters allocated to each layer. We evaluate the performance of dynamic reallocation methods in training deep convolutional networks and show that our method outperforms previous static and dynamic reparameterization methods, yielding the best accuracy for a fixed parameter budget, on par with accuracies obtained by iteratively pruning a pre-trained dense model. We further investigated the mechanisms underlying the superior generalization performance of the resultant sparse networks. We found that neither the structure, nor the initialization of the non-zero parameters were sufficient to explain the superior performance. Rather, effective learning crucially depended on the continuous exploration of the sparse network structure space during training. Our work suggests that exploring structural degrees of freedom during training is more effective than adding extra parameters to the network.

****

A Dynamical Systems Perspective on Nesterov Acceleration

Michael Muehlebach, Michael Jordan

We present a dynamical system framework for understanding Nesterov’s accelerated gradient method. In contrast to earlier work, our derivation does not rely on a vanishing step size argument. We show that Nesterov acceleration arises from discretizing an ordinary differential equation with a semi-implicit Euler integration scheme. We analyze both the underlying differential equation as well as the discretization to obtain insights into the phenomenon of acceleration. The analysis suggests that a curvature-dependent damping term lies at the heart of the phenomenon. We further establish connections between the discretized and the continuous-time dynamics.

****

Relational Pooling for Graph Representations

Ryan Murphy, Balasubramaniam Srinivasan, Vinayak Rao, Bruno Ribeiro

This work generalizes graph neural networks (GNNs) beyond those based on the Weisfeiler-Lehman (WL) algorithm, graph Laplacians, and diffusions. Our approach, denoted Relational Pooling (RP), draws from the theory of finite partial exchangeability to provide a framework with maximal representation power for graphs. RP can work with existing graph representation models and, somewhat counterintuitively, can make them even more powerful than the original WL isomorphism test. Additionally, RP allows architectures like Recurrent Neural Networks and Convolutional Neural Networks to be used in a theoretically sound approach for graph classification. We demonstrate improved performance of RP-based graph representations over state-of-the-art methods on a number of tasks.

****

Learning Optimal Fair Policies

Razieh Nabi, Daniel Malinsky, Ilya Shpitser

Systematic discriminatory biases present in our society influence the way data is collected and stored, the way variables are defined, and the way scientific findings are put into practice as policy. Automated decision procedures and learning algorithms applied to such data may serve to perpetuate existing injustice or unfairness in our society. In this paper, we consider how to make optimal but fair decisions, which “break the cycle of injustice” by correcting for the unfair dependence of both decisions and outcomes on sensitive features (e.g., variables that correspond to gender, race, disability, or other protected attributes). We use methods from causal inference and constrained optimization to learn optimal policies in a way that addresses multiple potential biases which afflict data analysis in sensitive contexts, extending the approach of Nabi & Shpitser (2018). Our proposal comes equipped with the theoretical guarantee that the chosen fair policy will induce a joint distribution for new instances that satisfies given fairness constraints. We illustrate our approach with both synthetic data and real criminal justice data.
****
Lexicographic and Depth-Sensitive Margins in Homogeneous and Non-Homogeneous Deep Models
Mor Shpigel Nacson, Suriya Gunasekar, Jason Lee, Nathan Srebro, Daniel Soudry
With an eye toward understanding complexity control in deep learning, we study how infinitesimal regularization or gradient descent optimization lead to margin maximizing solutions in both homogeneous and non homogeneous models, extending previous work that focused on infinitesimal regularization only in homogeneous models. To this end we study the limit of loss minimization with a diverging norm constraint (the “constrained path”), relate it to the limit of a “margin path” and characterize the resulting solution. For non-homogeneous ensemble models, which output is a sum of homogeneous sub-models, we show that this solution discards the shallowest sub-models if they are unnecessary. For homogeneous models, we show convergence to a “lexicographic max-margin solution”, and provide conditions under which max-margin solutions are also attained as the limit of unconstrained gradient descent.
****
A Wrapped Normal Distribution on Hyperbolic Space for Gradient-Based Learning
Yoshihiro Nagano, Shoichiro Yamaguchi, Yasuhiro Fujita, Masanori Koyama
Hyperbolic space is a geometry that is known to be well-suited for representation learning of data with an underlying hierarchical structure. In this paper, we present a novel hyperbolic distribution called hyperbolic wrapped distribution, a wrapped normal distribution on hyperbolic space whose density can be evaluated analytically and differentiated with respect to the parameters. Our distribution enables the gradient-based learning of the probabilistic models on hyperbolic space that could never have been considered before. Also, we can sample from this hyperbolic probability distribution without resorting to auxiliary means like rejection sampling. As applications of our distribution, we develop a hyperbolic-analog of variational autoencoder and a method of probabilistic word embedding on hyperbolic space. We demonstrate the efficacy of our distribution on various datasets including MNIST, Atari 2600 Breakout, and WordNet.
****
SGD without Replacement: Sharper Rates for General Smooth Convex Functions
Dheeraj Nagaraj, Prateek Jain, Praneeth Netrapalli
We study stochastic gradient descent without replacement (SGDo) for smooth convex functions. SGDo is widely observed to converge faster than true SGD where each sample is drawn independently with replacement (Bottou,2009) and hence, is more popular in practice. But it’s convergence properties are not well understood as sampling without replacement leads to coupling between iterates and gradients. By using method of exchangeable pairs to bound Wasserstein distance, we provide the first non-asymptotic results for SGDo when applied to general smooth, strongly-convex functions. In particular, we show that SGDo converges at a rate of $O(1/K^2)$ while SGD is known to converge at $O(1/K)$ rate, where K denotes the number of passes over data and is required to be large enough. Existing results for SGDo in this setting require additional Hessian Lipschitz assumption (Gurbuzb

alaban et al, 2015; HaoChen and Sra 2018). For small K, we show SGDo can achieve same convergence rate as SGD for general smooth strongly-convex functions. Existing results in this setting require $K=1$ and hold only for generalized linear models (Shamir,2016). In addition, by careful analysis of the coupling, for both large and small K, we obtain better dependence on problem dependent parameters like condition number.
****
Dropout as a Structured Shrinkage Prior
Eric Nalisnick, Jose Miguel Hernandez-Lobato, Padhraic Smyth
Dropout regularization of deep neural networks has been a mysterious yet effective tool to prevent overfitting. Explanations for its success range from the prevention of "co-adapted" weights to it being a form of cheap Bayesian inference. We propose a novel framework for understanding multiplicative noise in neural networks, considering continuous distributions as well as Bernoulli noise (i.e. dropout). We show that multiplicative noise induces structured shrinkage priors on a network’s weights. We derive the equivalence through reparametrization properties of scale mixtures and without invoking any approximations. Given the equivalence, we then show that dropout’s Monte Carlo training objective approximates marginal MAP estimation. We leverage these insights to propose a novel shrinkage framework for resnets, terming the prior ’automatic depth determination’ as it is the natural analog of automatic relevance determination for network depth. Lastly, we investigate two inference strategies that improve upon the aforementioned MAP approximation in regression benchmarks.
****
Hybrid Models with Deep and Invertible Features
Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, Balaji Lakshminarayanan
We propose a neural hybrid model consisting of a linear model defined on a set of features computed by a deep, invertible transformation (i.e. a normalizing flow). An attractive property of our model is that both p(features), the density of the features, and p(targets|features), the predictive distribution, can be computed exactly in a single feed-forward pass. We show that our hybrid model, despite the invertibility constraints, achieves similar accuracy to purely predictive models. Yet the generative component remains a good model of the input features despite the hybrid optimization objective. This offers additional capabilities such as detection of out-of-distribution inputs and enabling semi-supervised learning. The availability of the exact joint density p(targets, features) also allows us to compute many quantities readily, making our hybrid model a useful building block for downstream applications of probabilistic deep learning.
****
Learning Context-dependent Label Permutations for Multi-label Classification
Jinseok Nam, Young-Bum Kim, Eneldo Loza Mencia, Sunghyun Park, Ruhi Sarikaya, Johannes Fürnkranz
A key problem in multi-label classification is to utilize dependencies among the labels. Chaining classifiers are a simple technique for addressing this problem but current algorithms all assume a fixed, static label ordering. In this work, we propose a multi-label classification approach which allows to choose a dynamic, context-dependent label ordering. Our proposed approach consists of two sub-components: a simple EM-like algorithm which bootstraps the learned model, and a more elaborate approach based on reinforcement learning. Our experiments on three public multi-label classification benchmarks show that our proposed dynamic label ordering approach based on reinforcement learning outperforms recurrent neural networks with fixed label ordering across both bipartition and ranking measures on all the three datasets. As a result, we obtain a powerful sequence prediction-based algorithm for multi-label classification, which is able to efficiently and explicitly exploit label dependencies.
****
Zero-Shot Knowledge Distillation in Deep Networks
Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, Anirban Chakraborty

Knowledge distillation deals with the problem of training a smaller model (Stude
nt) from a high capacity source model (Teacher) so as to retain most of its perf
ormance. Existing approaches use either the training data or meta-data extracted
 from it in order to train the Student. However, accessing the dataset on which
the Teacher has been trained may not always be feasible if the dataset is very l
arge or it poses privacy or safety concerns (e.g., bio-metric or medical data).
Hence, in this paper, we propose a novel data-free method to train the Student f
rom the Teacher. Without even using any meta-data, we synthesize the Data Impres
sions from the complex Teacher model and utilize these as surrogates for the ori
ginal training data samples to transfer its learning to Student via knowledge di
stillation. We, therefore, dub our method “Zero-Shot Knowledge Distillation" and
 demonstrate that our framework results in competitive generalization performanc
e as achieved by distillation using the actual training data samples on multiple
 benchmark datasets.
****
A Framework for Bayesian Optimization in Embedded Subspaces
Amin Nayebi, Alexander Munteanu, Matthias Poloczek
We present a theoretically founded approach for high-dimensional Bayesian optimi
zation based on low-dimensional subspace embeddings. We prove that the error in
the Gaussian process model is bounded tightly when going from the original high-
dimensional search domain to the low-dimensional embedding. This implies that th
e optimization process in the low-dimensional embedding proceeds essentially as
if it were run directly on an unknown active subspace of low dimensionality. The
 argument applies to a large class of algorithms and GP models, including non-st
ationary kernels. Moreover, we provide an efficient implementation based on hash
ing and demonstrate empirically that this subspace embedding achieves considerab
ly better results than the previously proposed methods for high-dimensional BO b
ased on Gaussian matrix projections and structure-learning.
****
Phaseless PCA: Low-Rank Matrix Recovery from Column-wise Phaseless Measurements
Seyedehsara Nayer, Praneeth Narayanamurthy, Namrata Vaswani
This work proposes the first set of simple, practically useful, and provable alg
orithms for two inter-related problems. (i) The first is low-rank matrix recover
y from magnitude-only (phaseless) linear projections of each of its columns. Thi
s finds important applications in phaseless dynamic imaging, e.g., Fourier Ptych
ographic imaging of live biological specimens. Our guarantee shows that, in the
regime of small ranks, the sample complexity required is only a little larger th
an the order-optimal one, and much smaller than what standard (unstructured) pha
se retrieval methods need. %Moreover our algorithm is fast and memory-efficient
if only the minimum required number of measurements is used (ii) The second prob
lem we study is a dynamic extension of the above: it allows the low-dimensional
subspace from which each image/signal (each column of the low-rank matrix) is ge
nerated to change with time. We introduce a simple algorithm that is provably co
rrect as long as the subspace changes are piecewise constant.
****
Safe Grid Search with Optimal Complexity
Eugene Ndiaye, Tam Le, Olivier Fercoq, Joseph Salmon, Ichiro Takeuchi
Popular machine learning estimators involve regularization parameters that can b
e challenging to tune, and standard strategies rely on grid search for this task
. In this paper, we revisit the techniques of approximating the regularization p
ath up to predefined tolerance ϵ in a unified framework and show that i
ts complexity is $O(1/\sqrt[d]{\epsilon})$ for uniformly convex loss of order $d
 \geq 2$ and $O(1/\sqrt{\epsilon})$ for Generalized Self-Concordant functions. T
his framework encompasses least-squares but also logistic regression, a case tha
t as far as we know was not handled as precisely in previous works. We leverage
our technique to provide refined bounds on the validation error as well as a pra
ctical algorithm for hyperparameter tuning. The latter has global convergence gu
arantee when targeting a prescribed accuracy on the validation set. Last but not
 least, our approach helps relieving the practitioner from the (often neglected)
 task of selecting a stopping criterion when optimizing over the training set: o

ur method automatically calibrates this criterion based on the targeted accuracy on the validation set.
****
Learning to bid in revenue-maximizing auctions
Thomas Nedelec, Noureddine El Karoui, Vianney Perchet
We consider the problem of the optimization of bidding strategies in prior-dependent revenue-maximizing auctions, when the seller fixes the reserve prices based on the bid distributions. Our study is done in the setting where one bidder is strategic. Using a variational approach, we study the complexity of the original objective and we introduce a relaxation of the objective functional in order to use gradient descent methods. Our approach is simple, general and can be applied to various value distributions and revenue-maximizing mechanisms. The new strategies we derive yield massive uplifts compared to the traditional truthfully bidding strategy.
****
On Connected Sublevel Sets in Deep Learning
Quynh Nguyen
This paper shows that every sublevel set of the loss function of a class of deep over-parameterized neural nets with piecewise linear activation functions is connected and unbounded. This implies that the loss has no bad local valleys and all of its global minima are connected within a unique and potentially very large global valley.
****
Anomaly Detection With Multiple-Hypotheses Predictions
Duc Tam Nguyen, Zhongyu Lou, Michael Klar, Thomas Brox
In one-class-learning tasks, only the normal case (foreground) can be modeled with data, whereas the variation of all possible anomalies is too erratic to be described by samples. Thus, due to the lack of representative data, the wide-spread discriminative approaches cannot cover such learning tasks, and rather generative models,which attempt to learn the input density of the foreground, are used. However, generative models suffer from a large input dimensionality (as in images) and are typically inefficient learners.We propose to learn the data distribution of the foreground more efficiently with a multi-hypotheses autoencoder. Moreover, the model is criticized by a discriminator, which prevents artificial data modes not supported by data, and which enforces diversity across hypotheses. Our multiple-hypotheses-based anomaly detection framework allows the reliable identification of out-of-distribution samples. For anomaly detection on CIFAR-10, it yields up to 3.9% points improvement over previously reported results. On a real anomaly detection task, the approach reduces the error of the baseline models from 6.8% to 1.5%.
****
Non-Asymptotic Analysis of Fractional Langevin Monte Carlo for Non-Convex Optimization
Than Huy Nguyen, Umut Simsekli, Gael Richard
Recent studies on diffusion-based sampling methods have shown that Langevin Monte Carlo (LMC) algorithms can be beneficial for non-convex optimization, and rigorous theoretical guarantees have been proven for both asymptotic and finite-time regimes. Algorithmically, LMC-based algorithms resemble the well-known gradient descent (GD) algorithm, where the GD recursion is perturbed by an additive Gaussian noise whose variance has a particular form. Fractional Langevin Monte Carlo (FLMC) is a recently proposed extension of LMC, where the Gaussian noise is replaced by a heavy-tailed α-stable noise. As opposed to its Gaussian counterpart, these heavy-tailed perturbations can incur large jumps and it has been empirically demonstrated that the choice of α-stable noise can provide several advantages in modern machine learning problems, both in optimization and sampling contexts. However, as opposed to LMC, only asymptotic convergence properties of FLMC have been yet established. In this study, we analyze the non-asymptotic behavior of FLMC for non-convex optimization and prove finite-time bounds for its expected suboptimality. Our results show that the weak-error of FLMC increases faster than LMC, which suggests using smaller step-sizes in FLMC. We finally

extend our results to the case where the exact gradients are replaced by stochastic gradients and show that similar results hold in this setting as well.
****

Rotation Invariant Householder Parameterization for Bayesian PCA

Rajbir Nirwan, Nils Bertschinger

We consider probabilistic PCA and related factor models from a Bayesian perspective. These models are in general not identifiable as the likelihood has a rotational symmetry. This gives rise to complicated posterior distributions with continuous subspaces of equal density and thus hinders efficiency of inference as well as interpretation of obtained parameters. In particular, posterior averages over factor loadings become meaningless and only model predictions are unambiguous. Here, we propose a parameterization based on Householder transformations, which remove the rotational symmetry of the posterior. Furthermore, by relying on results from random matrix theory, we establish the parameter distribution which leaves the model unchanged compared to the original rotationally symmetric formulation. In particular, we avoid the need to compute the Jacobian determinant of the parameter transformation. This allows us to efficiently implement probabilistic PCA in a rotation invariant fashion in any state of the art toolbox. Here, we implemented our model in the probabilistic programming language Stan and illustrate it on several examples.
****

Lossless or Quantized Boosting with Integer Arithmetic

Richard Nock, Robert Williamson

In supervised learning, efficiency often starts with the choice of a good loss: support vector machines popularised Hinge loss, Adaboost popularised the exponential loss, etc. Recent trends in machine learning have highlighted the necessity for training routines to meet tight requirements on communication, bandwidth, energy, operations, encoding, among others. Fitting the often decades-old state of the art training routines into these new constraints does not go without pain and uncertainty or reduction in the original guarantees. Our paper starts with the design of a new strictly proper canonical, twice differentiable loss called the Q-loss. Importantly, its mirror update over (arbitrary) rational inputs uses only integer arithmetics – more precisely, the sole use of $+, -, /, \times, |.|$. We build a learning algorithm which is able, under mild assumptions, to achieve a lossless boosting-compliant training. We give conditions for a quantization of its main memory footprint, weights, to be done while keeping the whole algorithm boosting-compliant. Experiments display that the algorithm can achieve a fast convergence during the early boosting rounds compared to AdaBoost, even with a weight storage that can be 30+ times smaller. Lastly, we show that the Bayes risk of the Q-loss can be used as node splitting criterion for decision trees and guarantees optimal boosting convergence.
****

Training Neural Networks with Local Error Signals

Arild Nøkland, Lars Hiller Eidnes

Supervised training of neural networks for classification is typically performed with a global loss function. The loss function provides a gradient for the output layer, and this gradient is back-propagated to hidden layers to dictate an update direction for the weights. An alternative approach is to train the network with layer-wise loss functions. In this paper we demonstrate, for the first time, that layer-wise training can approach the state-of-the-art on a variety of image datasets. We use single-layer sub-networks and two different supervised loss functions to generate local error signals for the hidden layers, and we show that the combination of these losses help with optimization in the context of local learning. Using local errors could be a step towards more biologically plausible deep learning because the global error does not have to be transported back to hidden layers. A completely backprop free variant outperforms previously reported results among methods aiming for higher biological plausibility.
****

Remember and Forget for Experience Replay

Guido Novati, Petros Koumoutsakos

Experience replay (ER) is a fundamental component of off-policy deep reinforceme
nt learning (RL). ER recalls experiences from past iterations to compute gradien
t estimates for the current policy, increasing data-efficiency. However, the acc
uracy of such updates may deteriorate when the policy diverges from past behavio
rs and can undermine the performance of ER. Many algorithms mitigate this issue
by tuning hyper-parameters to slow down policy changes. An alternative is to act
ively enforce the similarity between policy and the experiences in the replay me
mory. We introduce Remember and Forget Experience Replay (ReF-ER), a novel metho
d that can enhance RL algorithms with parameterized policies. ReF-ER (1) skips g
radients computed from experiences that are too unlikely with the current policy
and (2) regulates policy changes within a trust region of the replayed behavior
s. We couple ReF-ER with Q-learning, deterministic policy gradient and off-polic
y gradient methods. We find that ReF-ER consistently improves the performance of
continuous-action, off-policy RL on fully observable benchmarks and partially o
bservable flow control problems.
****
Learning to Infer Program Sketches
Maxwell Nye, Luke Hewitt, Joshua Tenenbaum, Armando Solar-Lezama
Our goal is to build systems which write code automatically from the kinds of sp
ecifications humans can most easily provide, such as examples and natural langua
ge instruction. The key idea of this work is that a flexible combination of patt
ern recognition and explicit reasoning can be used to solve these complex progra
mming problems. We propose a method for dynamically integrating these types of i
nformation. Our novel intermediate representation and training algorithm allow a
program synthesis system to learn, without direct supervision, when to rely on
pattern recognition and when to perform symbolic search. Our model matches the m
emorization and generalization performance of neural synthesis and symbolic sear
ch, respectively, and achieves state-of-the-art performance on a dataset of simp
le English description-to-code programming problems.
****
Tensor Variable Elimination for Plated Factor Graphs
Fritz Obermeyer, Eli Bingham, Martin Jankowiak, Neeraj Pradhan, Justin Chiu, Ale
xander Rush, Noah Goodman
A wide class of machine learning algorithms can be reduced to variable eliminati
on on factor graphs. While factor graphs provide a unifying notation for these a
lgorithms, they do not provide a compact way to express repeated structure when
compared to plate diagrams for directed graphical models. To exploit efficient t
ensor algebra in graphs with plates of variables, we generalize undirected facto
r graphs to plated factor graphs and variable elimination to a tensor variable e
limination algorithm that operates directly on plated factor graphs. Moreover, w
e generalize complexity bounds based on treewidth and characterize the class of
plated factor graphs for which inference is tractable. As an application, we int
egrate tensor variable elimination into the Pyro probabilistic programming langu
age to enable exact inference in discrete latent variable models with repeated s
tructure. We validate our methods with experiments on both directed and undirect
ed graphical models, including applications to polyphonic music modeling, animal
movement modeling, and latent sentiment analysis.
****
Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models
Michael Oberst, David Sontag
We introduce an off-policy evaluation procedure for highlighting episodes where
applying a reinforcement learned (RL) policy is likely to have produced a substa
ntially different outcome than the observed policy. In particular, we introduce
a class of structural causal models (SCMs) for generating counterfactual traject
ories in finite partially observable Markov Decision Processes (POMDPs). We see
this as a useful procedure for off-policy “debugging” in high-risk settings (e.g
., healthcare); by decomposing the expected difference in reward between the RL
and observed policy into specific episodes, we can identify episodes where the c
ounterfactual difference in reward is most dramatic. This in turn can be used to
facilitate review of specific episodes by domain experts. We demonstrate the ut

ility of this procedure with a synthetic environment of sepsis management.
****
Model Function Based Conditional Gradient Method with Armijo-like Line Search
Peter Ochs, Yura Malitsky
The Conditional Gradient Method is generalized to a class of non-smooth non-conv ex optimization problems with many applications in machine learning. The propose d algorithm iterates by minimizing so-called model functions over the constraint set. Complemented with an Armijo line search procedure, we prove that subsequen ces converge to a stationary point. The abstract framework of model functions pr ovides great flexibility in the design of concrete algorithms. As special cases, for example, we develop an algorithm for additive composite problems and an alg orithm for non-linear composite problems which leads to a Gauss-Newton-type algo rithm. Both instances are novel in non-smooth non-convex optimization and come w ith numerous applications in machine learning. We perform an experiment on a non -linear robust regression problem and discuss the flexibility of the proposed fr amework in several matrix factorization formulations.
****
TensorFuzz: Debugging Neural Networks with Coverage-Guided Fuzzing
Augustus Odena, Catherine Olsson, David Andersen, Ian Goodfellow
Neural networks are difficult to interpret and debug. We introduce testing techn iques for neural networks that can discover errors occurring only for rare input s. Specifically, we develop coverage-guided fuzzing (CGF) methods for neural net works. In CGF, random mutations of inputs are guided by a coverage metric toward the goal of satisfying user-specified constraints. We describe how approximate nearest neighbor (ANN) algorithms can provide this coverage metric for neural ne tworks. We then combine these methods with techniques for property-based testing (PBT). In PBT, one asserts properties that a function should satisfy and the sy stem automatically generates tests exercising those properties. We then apply th is system to practical goals including (but not limited to) surfacing broken los s functions in popular GitHub repositories and making performance improvements t o TensorFlow. Finally, we release an open source library called TensorFuzz that implements the described techniques.
****
Scalable Learning in Reproducing Kernel Krein Spaces
Dino Oglic, Thomas Gärtner
We provide the first mathematically complete derivation of the Nystr{ö}m method for low-rank approximation of indefinite kernels and propose an efficient method for finding an approximate eigendecomposition of such kernel matrices. Building on this result, we devise highly scalable methods for learning in reproducing k ernel Krein spaces. The devised approaches provide a principled and theoreticall y well-founded means to tackle large scale learning problems with indefinite ker nels. The main motivation for our work comes from problems with structured repre sentations (e.g., graphs, strings, time-series), where it is relatively easy to devise a pairwise (dis)similarity function based on intuition and/or knowledge o f domain experts. Such functions are typically not positive definite and it is o ften well beyond the expertise of practitioners to verify this condition. The ef fectiveness of the devised approaches is evaluated empirically using indefinite kernels defined on structured and vectorial data representations.
****
Approximation and non-parametric estimation of ResNet-type convolutional neural networks
Kenta Oono, Taiji Suzuki
Convolutional neural networks (CNNs) have been shown to achieve optimal approxim ation and estimation error rates (in minimax sense) in several function classes. However, previous analyzed optimal CNNs are unrealistically wide and difficult to obtain via optimization due to sparse constraints in important function class es, including the Hölder class. We show a ResNet-type CNN can attain the minimax optimal error rates in these classes in more plausible situations – it can be d ense, and its width, channel size, and filter size are constant with respect to sample size. The key idea is that we can replicate the learning ability of Fully

-connected neural networks (FNNs) by tailored CNNs, as long as the FNNs have blo
ck-sparse structures. Our theory is general in a sense that we can automatically
 translate any approximation rate achieved by block-sparse FNNs into that by CNN
s. As an application, we derive approximation and estimation error rates of the
aformentioned type of CNNs for the Barron and Hölder classes with the same strat
egy.
****
Orthogonal Random Forest for Causal Inference
Miruna Oprescu, Vasilis Syrgkanis, Zhiwei Steven Wu
We propose the orthogonal random forest, an algorithm that combines Neyman-ortho
gonality to reduce sensitivity with respect to estimation error of nuisance para
meters with generalized random forests (Athey et al., 2017)—a flexible non-param
etric method for statistical estimation of conditional moment models using rando
m forests. We provide a consistency rate and establish asymptotic normality for
our estimator. We show that under mild assumptions on the consistency rate of th
e nuisance estimator, we can achieve the same error rate as an oracle with a pri
ori knowledge of these nuisance parameters. We show that when the nuisance funct
ions have a locally sparse parametrization, then a local ell_1-penalized regress
ion achieves the required rate. We apply our method to estimate heterogeneous tr
eatment effects from observational data with discrete treatments or continuous t
reatments, and we show that, unlike prior work, our method provably allows to co
ntrol for a high-dimensional set of variables under standard sparsity conditions
. We also provide a comprehensive empirical evaluation of our algorithm on both
synthetic and real data.
****
Inferring Heterogeneous Causal Effects in Presence of Spatial Confounding
Muhammad Osama, Dave Zachariah, Thomas B. Schön
We address the problem of inferring the causal effect of an exposure on an outco
me across space, using observational data. The data is possibly subject to unmea
sured confounding variables which, in a standard approach, must be adjusted for
by estimating a nuisance function. Here we develop a method that eliminates the
nuisance function, while mitigating the resulting errors-in-variables. The resul
t is a robust and accurate inference method for spatially varying heterogeneous
causal effects. The properties of the method are demonstrated on synthetic as we
ll as real data from Germany and the US.
****
Overparameterized Nonlinear Learning: Gradient Descent Takes the Shortest Path?
Samet Oymak, Mahdi Soltanolkotabi
Many modern learning tasks involve fitting nonlinear models which are trained in
 an overparameterized regime where the parameters of the model exceed the size o
f the training dataset. Due to this overparameterization, the training loss may
have infinitely many global minima and it is critical to understand the properti
es of the solutions found by first-order optimization schemes such as (stochasti
c) gradient descent starting from different initializations. In this paper we de
monstrate that when the loss has certain properties over a minimally small neigh
borhood of the initial point, first order methods such as (stochastic) gradient
descent have a few intriguing properties: (1) the iterates converge at a geometr
ic rate to a global optima even when the loss is nonconvex, (2) among all global
 optima of the loss the iterates converge to one with a near minimal distance to
 the initial point, (3) the iterates take a near direct route from the initial p
oint to this global optimum. As part of our proof technique, we introduce a new
potential function which captures the tradeoff between the loss function and the
 distance to the initial point as the iterations progress. The utility of our ge
neral theory is demonstrated for a variety of problem domains spanning low-rank
matrix recovery to shallow neural network training.
****
Multiplicative Weights Updates as a distributed constrained optimization algorit
hm: Convergence to second-order stationary points almost always
Ioannis Panageas, Georgios Piliouras, Xiao Wang
Non-concave maximization has been the subject of much recent study in the optimi

zation and machine learning communities, specifically in deep learning. Recent papers ([Ge et al. 2015, Lee et al 2017] and references therein) indicate that first order methods work well and avoid saddles points. Results as in [Lee \etal 2017], however, are limited to the unconstrained case or for cases where the critical points are in the interior of the feasibility set, which fail to capture some of the most interesting applications. In this paper we focus on constrained non-concave maximization. We analyze a variant of a well-established algorithm in machine learning called Multiplicative Weights Update (MWU) for the maximization problem $\max_{\mathbf{x} \in D} P(\mathbf{x})$, where P is non-concave, twice continuously differentiable and D is a product of simplices. We show that MWU converges almost always for small enough stepsizes to critical points that satisfy the second order KKT conditions, by combining techniques from dynamical systems as well as taking advantage of a recent connection between Baum Eagon inequality and MWU [Palaiopanos et al 2017].
****
Improving Adversarial Robustness via Promoting Ensemble Diversity
Tianyu Pang, Kun Xu, Chao Du, Ning Chen, Jun Zhu
Though deep neural networks have achieved significant progress on various tasks, often enhanced by model ensemble, existing high-performance models can be vulnerable to adversarial attacks. Many efforts have been devoted to enhancing the robustness of individual networks and then constructing a straightforward ensemble, e.g., by directly averaging the outputs, which ignores the interaction among networks. This paper presents a new method that explores the interaction among individual networks to improve robustness for ensemble models. Technically, we define a new notion of ensemble diversity in the adversarial setting as the diversity among non-maximal predictions of individual members, and present an adaptive diversity promoting (ADP) regularizer to encourage the diversity, which leads to globally better robustness for the ensemble by making adversarial examples difficult to transfer among individual members. Our method is computationally efficient and compatible with the defense methods acting on individual networks. Empirical results on various datasets verify that our method can improve adversarial robustness while maintaining state-of-the-art accuracy on normal examples.
****
Nonparametric Bayesian Deep Networks with Local Competition
Konstantinos Panousis, Sotirios Chatzis, Sergios Theodoridis
The aim of this work is to enable inference of deep networks that retain high accuracy for the least possible model complexity, with the latter deduced from the data during inference. To this end, we revisit deep networks that comprise competing linear units, as opposed to nonlinear units that do not entail any form of (local) competition. In this context, our main technical innovation consists in an inferential setup that leverages solid arguments from Bayesian nonparametrics. We infer both the needed set of connections or locally competing sets of units, as well as the required floating-point precision for storing the network parameters. Specifically, we introduce auxiliary discrete latent variables representing which initial network components are actually needed for modeling the data at hand, and perform Bayesian inference over them by imposing appropriate stick-breaking priors. As we experimentally show using benchmark datasets, our approach yields networks with less computational footprint than the state-of-the-art, and with no compromises in predictive accuracy.
****
Optimistic Policy Optimization via Multiple Importance Sampling
Matteo Papini, Alberto Maria Metelli, Lorenzo Lupo, Marcello Restelli
Policy Search (PS) is an effective approach to Reinforcement Learning (RL) for solving control tasks with continuous state-action spaces. In this paper, we address the exploration-exploitation trade-off in PS by proposing an approach based on Optimism in the Face of Uncertainty. We cast the PS problem as a suitable Multi Armed Bandit (MAB) problem, defined over the policy parameter space, and we propose a class of algorithms that effectively exploit the problem structure, by leveraging Multiple Importance Sampling to perform an off-policy estimation of the expected return. We show that the regret of the proposed approach is bounded

by $\widetilde{\mathcal{O}}(\sqrt{T})$ for both discrete and continuous paramete
r spaces. Finally, we evaluate our algorithms on tasks of varying difficulty, co
mparing them with existing MAB and RL algorithms.
****
Deep Residual Output Layers for Neural Language Generation
Nikolaos Pappas, James Henderson
Many tasks, including language generation, benefit from learning the structure o
f the output space, particularly when the space of output labels is large and th
e data is sparse. State-of-the-art neural language models indirectly capture the
 output space structure in their classifier weights since they lack parameter sh
aring across output labels. Learning shared output label mappings helps, but exi
sting methods have limited expressivity and are prone to overfitting. In this pa
per, we investigate the usefulness of more powerful shared mappings for output l
abels, and propose a deep residual output mapping with dropout between layers to
 better capture the structure of the output space and avoid overfitting. Evaluat
ions on three language generation tasks show that our output label mapping can m
atch or improve state-of-the-art recurrent and self-attention architectures, and
 suggest that the classifier does not necessarily need to be high-rank to better
 model natural language if it is better at capturing the structure of the output
 space.
****
Measurements of Three-Level Hierarchical Structure in the Outliers in the Spectr
um of Deepnet Hessians
Vardan Papyan
We expose a structure in deep classifying neural networks in the derivative of t
he logits with respect to the parameters of the model, which is used to explain
the existence of outliers in the spectrum of the Hessian. Previous works decompo
sed the Hessian into two components, attributing the outliers to one of them, th
e so-called Covariance of gradients. We show this term is not a Covariance but a
 second moment matrix, i.e., it is influenced by means of gradients. These means
 possess an additive two-way structure that is the source of the outliers in the
 spectrum. This structure can be used to approximate the principal subspace of t
he Hessian using certain "averaging" operations, avoiding the need for high-dime
nsional eigenanalysis. We corroborate this claim across different datasets, arch
itectures and sample sizes.
****
Generalized Majorization-Minimization
Sobhan Naderi Parizi, Kun He, Reza Aghajani, Stan Sclaroff, Pedro Felzenszwalb
Non-convex optimization is ubiquitous in machine learning. Majorization-Minimiza
tion (MM) is a powerful iterative procedure for optimizing non-convex functions
that works by optimizing a sequence of bounds on the function. In MM, the bound
at each iteration is required to touch the objective function at the optimizer o
f the previous bound. We show that this touching constraint is unnecessary and o
verly restrictive. We generalize MM by relaxing this constraint, and propose a n
ew optimization framework, named Generalized Majorization-Minimization (G-MM), t
hat is more flexible. For instance, G-MM can incorporate application-specific bi
ases into the optimization procedure without changing the objective function. We
 derive G-MM algorithms for several latent variable models and show empirically
that they consistently outperform their MM counterparts in optimizing non-convex
 objectives. In particular, G-MM algorithms appear to be less sensitive to initi
alization.
****
Variational Laplace Autoencoders
Yookoon Park, Chris Kim, Gunhee Kim
Variational autoencoders employ an amortized inference model to approximate the
posterior of latent variables. However, such amortized variational inference fac
es two challenges: (1) the limited posterior expressiveness of fully-factorized
Gaussian assumption and (2) the amortization error of the inference model. We pr
esent a novel approach that addresses both challenges. First, we focus on ReLU n
etworks with Gaussian output and illustrate their connection to probabilistic PC

A. Building on this observation, we derive an iterative algorithm that finds the mode of the posterior and apply fullcovariance Gaussian posterior approximation centered on the mode. Subsequently, we present a general framework named Variational Laplace Autoencoders (VLAEs) for training deep generative models. Based on the Laplace approximation of the latent variable posterior, VLAEs enhance the expressiveness of the posterior while reducing the amortization error. Empirical results on MNIST, Omniglot, Fashion-MNIST, SVHN and CIFAR10 show that the proposed approach significantly outperforms other recent amortized or iterative methods on the ReLU networks.
****

The Effect of Network Width on Stochastic Gradient Descent and Generalization: an Empirical Study

Daniel Park, Jascha Sohl-Dickstein, Quoc Le, Samuel Smith

We investigate how the final parameters found by stochastic gradient descent are influenced by over-parameterization. We generate families of models by increasing the number of channels in a base network, and then perform a large hyper-parameter search to study how the test error depends on learning rate, batch size, and network width. We find that the optimal SGD hyper-parameters are determined by a "normalized noise scale," which is a function of the batch size, learning rate, and initialization conditions. In the absence of batch normalization, the optimal normalized noise scale is directly proportional to width. Wider networks, with their higher optimal noise scale, also achieve higher test accuracy. These observations hold for MLPs, ConvNets, and ResNets, and for two different parameterization schemes ("Standard" and "NTK"). We observe a similar trend with batch normalization for ResNets. Surprisingly, since the largest stable learning rate is bounded, the largest batch size consistent with the optimal normalized noise scale decreases as the width increases.
****

Spectral Approximate Inference

Sejun Park, Eunho Yang, Se-Young Yun, Jinwoo Shin

Given a graphical model (GM), computing its partition function is the most essential inference task, but it is computationally intractable in general. To address the issue, iterative approximation algorithms exploring certain local structure/consistency of GM have been investigated as popular choices in practice. However, due to their local/iterative nature, they often output poor approximations or even do not converge, e.g., in low-temperature regimes (hard instances of large parameters). To overcome the limitation, we propose a novel approach utilizing the global spectral feature of GM. Our contribution is two-fold: (a) we first propose a fully polynomial-time approximation scheme (FPTAS) for approximating the partition function of GM associating with a low-rank coupling matrix; (b) for general high-rank GMs, we design a spectral mean-field scheme utilizing (a) as a subroutine, where it approximates a high-rank GM into a product of rank-1 GMs for an efficient approximation of the partition function. The proposed algorithm is more robust in its running time and accuracy than prior methods, i.e., neither suffers from the convergence issue nor depends on hard local structures, as demonstrated in our experiments.
****

Self-Supervised Exploration via Disagreement

Deepak Pathak, Dhiraj Gandhi, Abhinav Gupta

Efficient exploration is a long-standing problem in sensorimotor learning. Major advances have been demonstrated in noise-free, non-stochastic domains such as video games and simulation. However, most of these formulations either get stuck in environments with stochastic dynamics or are too inefficient to be scalable to real robotics setups. In this paper, we propose a formulation for exploration inspired by the work in active learning literature. Specifically, we train an ensemble of dynamics models and incentivize the agent to explore such that the disagreement of those ensembles is maximized. This allows the agent to learn skills by exploring in a self-supervised manner without any external reward. Notably, we further leverage the disagreement objective to optimize the agent’s policy in a differentiable manner, without using reinforcement learning, which results in

a sample-efficient exploration. We demonstrate the efficacy of this formulation across a variety of benchmark environments including stochastic-Atari, Mujoco and Unity. Finally, we implement our differentiable exploration on a real robot which learns to interact with objects completely from scratch. Project videos and code are at https://pathak22.github.io/exploration-by-disagreement/
****
Subspace Robust Wasserstein Distances
François-Pierre Paty, Marco Cuturi
Making sense of Wasserstein distances between discrete measures in high-dimensional settings remains a challenge. Recent work has advocated a two-step approach to improve robustness and facilitate the computation of optimal transport, using for instance projections on random real lines, or a preliminary quantization of the measures to reduce the size of their support. We propose in this work a “max-min” robust variant of the Wasserstein distance by considering the maximal possible distance that can be realized between two measures, assuming they can be projected orthogonally on a lower k-dimensional subspace. Alternatively, we show that the corresponding “min-max” OT problem has a tight convex relaxation which can be cast as that of finding an optimal transport plan with a low transportation cost, where the cost is alternatively defined as the sum of the k largest eigenvalues of the second order moment matrix of the displacements (or matchings) corresponding to that plan (the usual OT definition only considers the trace of that matrix). We show that both quantities inherit several favorable properties from the OT geometry. We propose two algorithms to compute the latter formulation using entropic regularization, and illustrate the interest of this approach empirically.
****
Fingerprint Policy Optimisation for Robust Reinforcement Learning
Supratik Paul, Michael A. Osborne, Shimon Whiteson
Policy gradient methods ignore the potential value of adjusting environment variables: unobservable state features that are randomly determined by the environment in a physical setting, but are controllable in a simulator. This can lead to slow learning, or convergence to suboptimal policies, if the environment variable has a large impact on the transition dynamics. In this paper, we present fingerprint policy optimisation (FPO), which finds a policy that is optimal in expectation across the distribution of environment variables. The central idea is to use Bayesian optimisation (BO) to actively select the distribution of the environment variable that maximises the improvement generated by each iteration of the policy gradient method. To make this BO practical, we contribute two easy-to-compute low-dimensional fingerprints of the current policy. Our experiments show that FPO can efficiently learn policies that are robust to significant rare events, which are unlikely to be observable under random sampling, but are key to learning good policies.
****
COMIC: Multi-view Clustering Without Parameter Selection
Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, Joey Tianyi Zhou
In this paper, we study two challenges in clustering analysis, namely, how to cluster multi-view data and how to perform clustering without parameter selection on cluster size. To this end, we propose a novel objective function to project raw data into one space in which the projection embraces the geometric consistency (GC) and the cluster assignment consistency (CAC). To be specific, the GC aims to learn a connection graph from a projection space wherein the data points are connected if and only if they belong to the same cluster. The CAC aims to minimize the discrepancy of pairwise connection graphs induced from different views based on the view-consensus assumption, i.e., different views could produce the same cluster assignment structure as they are different portraits of the same object. Thanks to the view-consensus derived from the connection graph, our method could achieve promising performance in learning view-specific representation and eliminating the heterogeneous gaps across different views. Furthermore, with the proposed objective, it could learn almost all parameters including the cluster number from data without labor-intensive parameter selection. Extensive experim

ental results show the promising performance achieved by our method on five data sets comparing with nine state-of-the-art multi-view clustering approaches.
****

Domain Agnostic Learning with Disentangled Representations
Xingchao Peng, Zijun Huang, Ximeng Sun, Kate Saenko
Unsupervised model transfer has the potential to greatly improve the generalizability of deep models to novel domains. Yet the current literature assumes that the separation of target data into distinct domains is known a priori. In this paper, we propose the task of Domain-Agnostic Learning (DAL): How to transfer knowledge from a labeled source domain to unlabeled data from arbitrary target domains? To tackle this problem, we devise a novel Deep Adversarial Disentangled Autoencoder (DADA) capable of disentangling domain-specific features from class identity. We demonstrate experimentally that when the target domain labels are unknown, DADA leads to state-of-the-art performance on several image classification datasets.
****

Collaborative Channel Pruning for Deep Networks
Hanyu Peng, Jiaxiang Wu, Shifeng Chen, Junzhou Huang
Deep networks have achieved impressive performance in various domains, but their applications are largely limited by the prohibitive computational overhead. In this paper, we propose a novel algorithm, namely collaborative channel pruning (CCP), to reduce the computational overhead with negligible performance degradation. The joint impact of pruned/preserved channels on the loss function is quantitatively analyzed, and such interchannel dependency is exploited to determine which channels to be pruned. The channel selection problem is then reformulated as a constrained 0-1 quadratic optimization problem, and the Hessian matrix, which is essential in constructing the above optimization, can be efficiently approximated. Empirical evaluation on two benchmark data sets indicates that our proposed CCP algorithm achieves higher classification accuracy with similar computational complexity than other stateof-the-art channel pruning algorithms
****

Exploiting structure of uncertainty for efficient matroid semi-bandits
Pierre Perrault, Vianney Perchet, Michal Valko
We improve the efficiency of algorithms for stochastic combinatorial semi-bandits. In most interesting problems, state-of-the-art algorithms take advantage of structural properties of rewards, such as independence. However, while being minimax optimal in terms of regret, these algorithms are intractable. In our paper, we first reduce their implementation to a specific submodular maximization. Then, in case of matroid constraints, we design adapted approximation routines, thereby providing the first efficient algorithms that exploit the reward structure. In particular, we improve the state-of-the-art efficient gap-free regret bound by a factor sqrt(k), where k is the maximum action size. Finally, we show how our improvement translates to more general budgeted combinatorial semi-bandits.
****

Cognitive model priors for predicting human decisions
David D. Bourgin, Joshua C. Peterson, Daniel Reichman, Stuart J. Russell, Thomas L. Griffiths
Human decision-making underlies all economic behavior. For the past four decades, human decision-making under uncertainty has continued to be explained by theoretical models based on prospect theory, a framework that was awarded the Nobel Prize in Economic Sciences. However, theoretical models of this kind have developed slowly, and robust, high-precision predictive models of human decisions remain a challenge. While machine learning is a natural candidate for solving these problems, it is currently unclear to what extent it can improve predictions obtained by current theories. We argue that this is mainly due to data scarcity, since noisy human behavior requires massive sample sizes to be accurately captured by off-the-shelf machine learning methods. To solve this problem, what is needed are machine learning models with appropriate inductive biases for capturing human behavior, and larger datasets. We offer two contributions towards this end: first, we construct “cognitive model priors” by pretraining neural networks with s

ynthetic data generated by cognitive models (i.e., theoretical models developed by cognitive psychologists). We find that fine-tuning these networks on small datasets of real human decisions results in unprecedented state-of-the-art improvements on two benchmark datasets. Second, we present the first large-scale dataset for human decision-making, containing over 240,000 human judgments across over 13,000 decision problems. This dataset reveals the circumstances where cognitive model priors are useful, and provides a new standard for benchmarking prediction of human decisions under uncertainty.
****
Towards Understanding Knowledge Distillation
Mary Phuong, Christoph Lampert
Knowledge distillation, i.e., one classifier being trained on the outputs of another classifier, is an empirically very successful technique for knowledge transfer between classifiers. It has even been observed that classifiers learn much faster and more reliably if trained with the outputs of another classifier as soft labels, instead of from ground truth data. So far, however, there is no satisfactory theoretical explanation of this phenomenon. In this work, we provide the first insights into the working mechanisms of distillation by studying the special case of linear and deep linear classifiers. Specifically, we prove a generalization bound that establishes fast convergence of the expected risk of a distillation-trained linear classifier. From the bound and its proof we extract three key factors that determine the success of distillation: * data geometry – geometric properties of the data distribution, in particular class separation, has a direct influence on the convergence speed of the risk; * optimization bias – gradient descent optimization finds a very favorable minimum of the distillation objective; and * strong monotonicity – the expected risk of the student classifier always decreases when the size of the training set grows.
****
Temporal Gaussian Mixture Layer for Videos
Aj Piergiovanni, Michael Ryoo
We introduce a new convolutional layer named the Temporal Gaussian Mixture (TGM) layer and present how it can be used to efficiently capture longer-term temporal information in continuous activity videos. The TGM layer is a temporal convolutional layer governed by a much smaller set of parameters (e.g., location/variance of Gaussians) that are fully differentiable. We present our fully convolutional video models with multiple TGM layers for activity detection. The extensive experiments on multiple datasets, including Charades and MultiTHUMOS, confirm the effectiveness of TGM layers, significantly outperforming the state-of-the-arts.
****
Voronoi Boundary Classification: A High-Dimensional Geometric Approach via Weighted Monte Carlo Integration
Vladislav Polianskii, Florian T. Pokorny
Voronoi cell decompositions provide a classical avenue to classification. Typical approaches however only utilize point-wise cell-membership information by means of nearest neighbor queries and do not utilize further geometric information about Voronoi cells since the computation of Voronoi diagrams is prohibitively expensive in high dimensions. We propose a Monte-Carlo integration based approach that instead computes a weighted integral over the boundaries of Voronoi cells, thus incorporating additional information about the Voronoi cell structure. We demonstrate the scalability of our approach in up to 3072 dimensional spaces and analyze convergence based on the number of Monte Carlo samples and choice of weight functions. Experiments comparing our approach to Nearest Neighbors, SVM and Random Forests indicate that while our approach performs similarly to Random Forests for large data sizes, the algorithm exhibits non-trivial data-dependent performance characteristics for smaller datasets and can be analyzed in terms of a geometric confidence measure, thus adding to the repertoire of geometric approaches to classification while having the benefit of not requiring any model changes or retraining as new training samples or classes are added.
****
On Variational Bounds of Mutual Information

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, George Tucker

Estimating and optimizing Mutual Information (MI) is core to many problems in machine learning, but bounding MI in high dimensions is challenging. To establish tractable and scalable objectives, recent work has turned to variational bounds parameterized by neural networks. However, the relationships and tradeoffs between these bounds remains unclear. In this work, we unify these recent developments in a single framework. We find that the existing variational lower bounds degrade when the MI is large, exhibiting either high bias or high variance. To address this problem, we introduce a continuum of lower bounds that encompasses previous bounds and flexibly trades off bias and variance. On high-dimensional, controlled problems, we empirically characterize the bias and variance of the bounds and their gradients and demonstrate the effectiveness of these new bounds for estimation and representation learning.

****

Hiring Under Uncertainty

Manish Purohit, Sreenivas Gollapudi, Manish Raghavan

In this paper we introduce the hiring under uncertainty problem to model the questions faced by hiring committees in large enterprises and universities alike. Given a set of n eligible candidates, the decision maker needs to choose the sequence of candidates to make offers so as to hire the k best candidates. However, candidates may choose to reject an offer (for instance, due to a competing offer) and the decision maker has a time limit by which all positions must be filled. Given an estimate of the probabilities of acceptance for each candidate, the hiring under uncertainty problem is to design a strategy of making offers so that the total expected value of all candidates hired by the time limit is maximized. We provide a 2-approximation algorithm for the setting where offers must be made in sequence, an 8-approximation when offers may be made in parallel, and a 10-approximation for the more general stochastic knapsack setting with finite probes.

****

SAGA with Arbitrary Sampling

Xun Qian, Zheng Qu, Peter Richtárik

We study the problem of minimizing the average of a very large number of smooth functions, which is of key importance in training supervised learning models. One of the most celebrated methods in this context is the SAGA algorithm of Defazio et al. (2014). Despite years of research on the topic, a general-purpose version of SAGA—one that would include arbitrary importance sampling and minibatching schemes—does not exist. We remedy this situation and propose a general and flexible variant of SAGA following the arbitrary sampling paradigm. We perform an iteration complexity analysis of the method, largely possible due to the construction of new stochastic Lyapunov functions. We establish linear convergence rates in the smooth and strongly convex regime, and under certain error bound conditions also in a regime without strong convexity. Our rates match those of the primal-dual method Quartz (Qu et al., 2015) for which an arbitrary sampling analysis is available, which makes a significant step towards closing the gap in our understanding of complexity of primal and dual methods for finite sum problems. Finally, we show through experiments that specific variants of our general SAGA method can perform better in practice than other competing methods.

****

SGD: General Analysis and Improved Rates

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, Peter Richtárik

We propose a general yet simple theorem describing the convergence of SGD under the arbitrary sampling paradigm. Our theorem describes the convergence of an infinite array of variants of SGD, each of which is associated with a specific probability law governing the data selection rule used to form minibatches. This is the first time such an analysis is performed, and most of our variants of SGD were never explicitly considered in the literature before. Our analysis relies on the recently introduced notion of expected smoothness and does not rely on a uniform bound on the variance of the stochastic gradients. By specializing our theo

rem to different mini-batching strategies, such as sampling with replacement and independent sampling, we derive exact expressions for the stepsize as a function of the mini-batch size. With this we can also determine the mini-batch size that optimizes the total complexity, and show explicitly that as the variance of the stochastic gradient evaluated at the minimum grows, so does the optimal mini-batch size. For zero variance, the optimal mini-batch size is one. Moreover, we prove insightful stepsize-switching rules which describe when one should switch from a constant to a decreasing stepsize regime.
****
AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss
Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Mark Hasegawa-Johnson
Despite the progress in voice conversion, many-to-many voice conversion trained on non-parallel data, as well as zero-shot voice conversion, remains under-explored. Deep style transfer algorithms, generative adversarial networks (GAN) in particular, are being applied as new solutions in this field. However, GAN training is very sophisticated and difficult, and there is no strong evidence that its generated speech is of good perceptual quality. In this paper, we propose a new style transfer scheme that involves only an autoencoder with a carefully designed bottleneck. We formally show that this scheme can achieve distribution-matching style transfer by training only on self-reconstruction loss. Based on this scheme, we proposed AutoVC, which achieves state-of-the-art results in many-to-many voice conversion with non-parallel data, and which is the first to perform zero-shot voice conversion.
****
Fault Tolerance in Iterative-Convergent Machine Learning
Aurick Qiao, Bryon Aragam, Bingjing Zhang, Eric Xing
Machine learning (ML) training algorithms often possess an inherent self-correcting behavior due to their iterative- convergent nature. Recent systems exploit this property to achieve adaptability and efficiency in unreliable computing environments by relaxing the consistency of execution and allowing calculation errors to be self-corrected during training. However, the behavior of such systems are only well understood for specific types of calculation errors, such as those caused by staleness, reduced precision, or asynchronicity, and for specific algorithms, such as stochastic gradient descent. In this paper, we develop a general framework to quantify the effects of calculation errors on iterative-convergent algorithms. We then use this framework to derive a worst-case upper bound on the cost of arbitrary perturbations to model parameters during training and to design new strategies for checkpoint-based fault tolerance. Our system, SCAR, can reduce the cost of partial failures by 78%{–}95% when compared with traditional checkpoint-based fault tolerance across a variety of ML models and training algorithms, providing near-optimal performance in recovering from failures.
****
Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition
Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, Colin Raffel
Adversarial examples are inputs to machine learning models designed by an adversary to cause an incorrect output. So far, adversarial examples have been studied most extensively in the image domain. In this domain, adversarial examples can be constructed by imperceptibly modifying images to cause misclassification, and are practical in the physical world. In contrast, current targeted adversarial examples on speech recognition systems have neither of these properties: humans can easily identify the adversarial perturbations, and they are not effective when played over-the-air. This paper makes progress on both of these fronts. First, we develop effectively imperceptible audio adversarial examples (verified through a human study) by leveraging the psychoacoustic principle of auditory masking, while retaining 100% targeted success rate on arbitrary full-sentence targets. Then, we make progress towards physical-world audio adversarial examples by constructing perturbations which remain effective even after applying highly-realistic simulated environmental distortions.
****

GMNN: Graph Markov Neural Networks
Meng Qu, Yoshua Bengio, Jian Tang
This paper studies semi-supervised object classification in relational data, which is a fundamental problem in relational data modeling. The problem has been extensively studied in the literature of both statistical relational learning (e.g. relational Markov networks) and graph neural networks (e.g. graph convolutional networks). Statistical relational learning methods can effectively model the dependency of object labels through conditional random fields for collective classification, whereas graph neural networks learn effective object representations for classification through end-to-end training. In this paper, we propose the Graph Markov Neural Network (GMNN) that combines the advantages of both worlds. A GMNN models the joint distribution of object labels with a conditional random field, which can be effectively trained with the variational EM algorithm. In the E-step, one graph neural network learns effective object representations for approximating the posterior distributions of object labels. In the M-step, another graph neural network is used to model the local label dependency. Experiments on object classification, link classification, and unsupervised node representation learning show that GMNN achieves state-of-the-art results.
****
Nonlinear Distributional Gradient Temporal-Difference Learning
Chao Qu, Shie Mannor, Huan Xu
We devise a distributional variant of gradient temporal-difference (TD) learning. Distributional reinforcement learning has been demonstrated to outperform the regular one in the recent study \citep{bellemare2017distributional}. In the policy evaluation setting, we design two new algorithms called distributional GTD2 and distributional TDC using the Cram{é}r distance on the distributional version of the Bellman error objective function, which inherits advantages of both the nonlinear gradient TD algorithms and the distributional RL approach. In the control setting, we propose the distributional Greedy-GQ using similar derivation. We prove the asymptotic almost-sure convergence of distributional GTD2 and TDC to a local optimal solution for general smooth function approximators, which includes neural networks that have been widely used in recent study to solve the real-life RL problems. In each step, the computational complexity of above three algorithms is linear w.r.t. the number of the parameters of the function approximator, thus can be implemented efficiently for neural networks.
****
Learning to Collaborate in Markov Decision Processes
Goran Radanovic, Rati Devidze, David Parkes, Adish Singla
We consider a two-agent MDP framework where agents repeatedly solve a task in a collaborative setting. We study the problem of designing a learning algorithm for the first agent (A1) that facilitates a successful collaboration even in cases when the second agent (A2) is adapting its policy in an unknown way. The key challenge in our setting is that the first agent faces non-stationarity in rewards and transitions because of the adaptive behavior of the second agent. We design novel online learning algorithms for agent A1 whose regret decays as $O(T^{1-\frac{3}{7} \cdot \alpha})$ with T learning episodes provided that the magnitude of agent A2’s policy changes between any two consecutive episodes are upper bounded by $O(T^{-\alpha})$. Here, the parameter α is assumed to be strictly greater than 0, and we show that this assumption is necessary provided that the learning parity with noise problem is computationally hard. We show that sub-linear regret of agent A1 further implies near-optimality of the agents’ joint return for MDPs that manifest the properties of a smooth game.
****
Meta-Learning Neural Bloom Filters
Jack Rae, Sergey Bartunov, Timothy Lillicrap
There has been a recent trend in training neural networks to replace data structures that have been crafted by hand, with an aim for faster execution, better accuracy, or greater compression. In this setting, a neural data structure is instantiated by training a network over many epochs of its inputs until convergence. In applications where inputs arrive at high throughput, or are ephemeral, train

ing a network from scratch is not practical. This motivates the need for few-shot neural data structures. In this paper we explore the learning of approximate set membership over a set of data in one-shot via meta-learning. We propose a novel memory architecture, the Neural Bloom Filter, which is able to achieve significant compression gains over classical Bloom Filters and existing memory-augmented neural networks.
****
Direct Uncertainty Prediction for Medical Second Opinions
Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Bobby Kleinberg, Sendhil Mullainathan, Jon Kleinberg
The issue of disagreements amongst human experts is a ubiquitous one in both machine learning and medicine. In medicine, this often corresponds to doctor disagreements on a patient diagnosis. In this work, we show that machine learning models can be successfully trained to give uncertainty scores to data instances that result in high expert disagreements. In particular, they can identify patient cases that would benefit most from a medical second opinion. Our central methodological finding is that Direct Uncertainty Prediction (DUP), training a model to predict an uncertainty score directly from the raw patient features, works better than Uncertainty Via Classification, the two step process of training a classifier and postprocessing the output distribution to give an uncertainty score. We show this both with a theoretical result, and on extensive evaluations on a large scale medical imaging application.
****
Game Theoretic Optimization via Gradient-based Nikaido-Isoda Function
Arvind Raghunathan, Anoop Cherian, Devesh Jha
Computing Nash equilibrium (NE) of multi-player games has witnessed renewed interest due to recent advances in generative adversarial networks. However, computing equilibrium efficiently is challenging. To this end, we introduce the Gradient-based Nikaido-Isoda (GNI) function which serves: (i) as a merit function, vanishing only at the first-order stationary points of each player’s optimization problem, and (ii) provides error bounds to a stationary Nash point. Gradient descent is shown to converge sublinearly to a first-order stationary point of the GNI function. For the particular case of bilinear min-max games and multi-player quadratic games, the GNI function is convex. Hence, the application of gradient descent in this case yields linear convergence to an NE (when one exists). In our numerical experiments, we observe that the GNI formulation always converges to the first-order stationary point of each player’s optimization problem.
****
On the Spectral Bias of Neural Networks
Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, Aaron Courville
Neural networks are known to be a class of highly expressive functions able to fit even random input-output mappings with 100% accuracy. In this work we present properties of neural networks that complement this aspect of expressivity. By using tools from Fourier analysis, we highlight a learning bias of deep networks towards low frequency functions – i.e. functions that vary globally without local fluctuations – which manifests itself as a frequency-dependent learning speed. Intuitively, this property is in line with the observation that over-parameterized networks prioritize learning simple patterns that generalize across data samples. We also investigate the role of the shape of the data manifold by presenting empirical and theoretical evidence that, somewhat counter-intuitively, learning higher frequencies gets easier with increasing manifold complexity.
****
Look Ma, No Latent Variables: Accurate Cutset Networks via Compilation
Tahrima Rahman, Shasha Jin, Vibhav Gogate
Tractable probabilistic models obviate the need for unreliable approximate inference approaches and as a result often yield accurate query answers in practice. However, most tractable models that achieve state-of-the-art generalization performance (measured using test set likelihood score) use latent variables. Such models admit poly-time marginal (MAR) inference but do not admit poly-time (full)

maximum-a-posteriori (MAP) inference. To address this problem, in this paper, we propose a novel approach for inducing cutset networks, a well-known tractable, highly interpretable representation that does not use latent variables and admits linear time MAR as well as MAP inference. Our approach addresses a major limitation of existing techniques that learn cutset networks from data in that their accuracy is quite low as compared to latent variable models such as ensembles of cutset networks and sum-product networks. The key idea in our approach is to construct deep cutset networks by not only learning them from data but also compiling them from a more accurate latent tractable model. We show experimentally that our new approach yields more accurate MAP estimates as compared with existing approaches and significantly improves the test set log-likelihood score of cutset networks bringing them closer in terms of generalization performance to latent variable models.

****

Does Data Augmentation Lead to Positive Margin?

Shashank Rajput, Zhili Feng, Zachary Charles, Po-Ling Loh, Dimitris Papailiopoulos

Data augmentation (DA) is commonly used during model training, as it significantly improves test error and model robustness. DA artificially expands the training set by applying random noise, rotations, crops, or even adversarial perturbations to the input data. Although DA is widely used, its capacity to provably improve robustness is not fully understood. In this work, we analyze the robustness that DA begets by quantifying the margin that DA enforces on empirical risk minimizers. We first focus on linear separators, and then a class of nonlinear models whose labeling is constant within small convex hulls of data points. We present lower bounds on the number of augmented data points required for non-zero margin, and show that commonly used DA techniques may only introduce significant margin after adding exponentially many points to the data set.

****

Efficient Off-Policy Meta-Reinforcement Learning via Probabilistic Context Variables

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, Deirdre Quillen

Deep reinforcement learning algorithms require large amounts of experience to learn an individual task. While meta-reinforcement learning (meta-RL) algorithms can enable agents to learn new skills from small amounts of experience, several major challenges preclude their practicality. Current methods rely heavily on on-policy experience, limiting their sample efficiency. They also lack mechanisms to reason about task uncertainty when adapting to new tasks, limiting their effectiveness on sparse reward problems. In this paper, we address these challenges by developing an off-policy meta-RL algorithm that disentangles task inference and control. In our approach, we perform online probabilistic filtering of latent task variables to infer how to solve a new task from small amounts of experience. This probabilistic interpretation enables posterior sampling for structured and efficient exploration. We demonstrate how to integrate these task variables with off-policy RL algorithms to achieve both meta-training and adaptation efficiency. Our method outperforms prior algorithms in sample efficiency by 20-100X as well as in asymptotic performance on several meta-RL benchmarks.

****

Screening rules for Lasso with non-convex Sparse Regularizers

Alain Rakotomamonjy, Gilles Gasso, Joseph Salmon

Leveraging on the convexity of the Lasso problem, screening rules help in accelerating solvers by discarding irrelevant variables, during the optimization process. However, because they provide better theoretical guarantees in identifying relevant■variables, several non-convex regularizers for the Lasso have been proposed in the literature. This work is the first that introduces a screening rule strategy into a non-convex Lasso solver. The approach we propose is based on a iterative majorization-minimization (MM) strategy that includes a screening rule in the inner solver and a condition for propagating screened variables between iterations of MM. In addition to improve efficiency of solvers, we also provide guarantees that the inner solver is able to identify the zeros components of its c

ritical point in finite time. Our experimental analysis illustrates the significant computational gain brought by the new screening rule compared to classical coordinate-descent or proximal gradient descent methods.
****

Topological Data Analysis of Decision Boundaries with Application to Model Selection

Karthikeyan Natesan Ramamurthy, Kush Varshney, Krishnan Mody

We propose the labeled Cech complex, the plain labeled Vietoris-Rips complex, and the locally scaled labeled Vietoris-Rips complex to perform persistent homology inference of decision boundaries in classification tasks. We provide theoretical conditions and analysis for recovering the homology of a decision boundary from samples. Our main objective is quantification of deep neural network complexity to enable matching of datasets to pre-trained models to facilitate the functioning of AI marketplaces; we report results for experiments using MNIST, Fashion MNIST, and CIFAR10.
****

HyperGAN: A Generative Model for Diverse, Performant Neural Networks

Neale Ratzlaff, Li Fuxin

We introduce HyperGAN, a generative model that learns to generate all the parameters of a deep neural network. HyperGAN first transforms low dimensional noise into a latent space, which can be sampled from to obtain diverse, performant sets of parameters for a target architecture. We utilize an architecture that bears resemblance to generative adversarial networks, but we evaluate the likelihood of generated samples with a classification loss. This is equivalent to minimizing the KL-divergence between the distribution of generated parameters, and the unknown true parameter distribution. We apply HyperGAN to classification, showing that HyperGAN can learn to generate parameters which solve the MNIST and CIFAR-10 datasets with competitive performance to fully supervised learning, while also generating a rich distribution of effective parameters. We also show that HyperGAN can also provide better uncertainty estimates than standard ensembles. This is evidenced by the ability of HyperGAN-generated ensembles to detect out of distribution data as well as adversarial examples.
****

Efficient On-Device Models using Neural Projections

Sujith Ravi

Many applications involving visual and language understanding can be effectively solved using deep neural networks. Even though these techniques achieve state-of-the-art results, it is very challenging to apply them on devices with limited memory and computational capacity such as mobile phones, smart watches and IoT. We propose a neural projection approach for training compact on-device neural networks. We introduce "projection" networks that use locality-sensitive projections to generate compact binary representations and learn small neural networks with computationally efficient operations. We design a joint optimization framework where the projection network can be trained from scratch or leverage existing larger neural networks such as feed-forward NNs, CNNs or RNNs. The trained neural projection network can be directly used for inference on device at low memory and computation cost. We demonstrate the effectiveness of this as a general-purpose approach for significantly shrinking memory requirements of different types of neural networks while preserving good accuracy on multiple visual and text classification tasks.
****

A Block Coordinate Descent Proximal Method for Simultaneous Filtering and Parameter Estimation

Ramin Raziperchikolaei, Harish Bhat

We propose and analyze a block coordinate descent proximal algorithm (BCD-prox) for simultaneous filtering and parameter estimation of ODE models. As we show on ODE systems with up to d=40 dimensions, as compared to state-of-the-art methods, BCD-prox exhibits increased robustness (to noise, parameter initialization, and hyperparameters), decreased training times, and improved accuracy of both filtered states and estimated parameters. We show how BCD-prox can be used with mult

istep numerical discretizations, and we establish convergence of BCD-prox under hypotheses that include real systems of interest.
****
Do ImageNet Classifiers Generalize to ImageNet?
Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, Vaishaal Shankar
We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of 3% - 15% on CIFAR-10 and 11% - 14% on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models' inability to generalize to slightly "harder" images than those found in the original test sets.
****
Fast Rates for a kNN Classifier Robust to Unknown Asymmetric Label Noise
Henry Reeve, Ata Kaban
We consider classification in the presence of class-dependent asymmetric label noise with unknown noise probabilities. In this setting, identifiability conditions are known, but additional assumptions were shown to be required for finite sample rates, and so far only the parametric rate has been obtained. Assuming these identifiability conditions, together with a measure-smoothness condition on the regression function and Tsybakov's margin condition, we show that the Robust kNN classifier of Gao et al. attains, the mini-max optimal rates of the noise-free setting, up to a log factor, even when trained on data with unknown asymmetric label noise. Hence, our results provide a solid theoretical backing for this empirically successful algorithm. By contrast the standard kNN is not even consistent in the setting of asymmetric label noise. A key idea in our analysis is a simple kNN based method for estimating the maximum of a function that requires far less assumptions than existing mode estimators do, and which may be of independent interest for noise proportion estimation and randomised optimisation problems.
****
Almost Unsupervised Text to Speech and Automatic Speech Recognition
Yi Ren, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu
Text to speech (TTS) and automatic speech recognition (ASR) are two dual tasks in speech processing and both achieve impressive performance thanks to the recent advance in deep learning and large amount of aligned speech and text data. However, the lack of aligned data poses a major practical problem for TTS and ASR on low-resource languages. In this paper, by leveraging the dual nature of the two tasks, we propose an almost unsupervised learning method that only leverages few hundreds of paired data and extra unpaired data for TTS and ASR. Our method consists of the following components: (1) denoising auto-encoder, which reconstructs speech and text sequences respectively to develop the capability of language modeling both in speech and text domain; (2) dual transformation, where the TTS model transforms the text y into speech $\hat{x}$, and the ASR model leverages the transformed pair $(\hat{x},y)$ for training, and vice versa, to boost the accuracy of the two tasks; (3) bidirectional sequence modeling, which address the error propagation problem especially in the long speech and text sequence when training with few paired data; (4) a unified model structure, which combines all the above components for TTS and ASR based on Transformer model. Our method achieves 99.84% in terms of word level intelligible rate and 2.68 MOS for TTS, and 11.7% PER for ASR on LJSpeech dataset, by leveraging only 200 paired speech and text data (about 20 minutes audio), together with extra unpaired speech and text data.
****
Adaptive Antithetic Sampling for Variance Reduction
Hongyu Ren, Shengjia Zhao, Stefano Ermon
Variance reduction is crucial in stochastic estimation and optimization problems

. Antithetic sampling reduces the variance of a Monte Carlo estimator by drawing correlated, rather than independent, samples. However, designing an effective correlation structure is challenging and application specific, thus limiting the practical applicability of these methods. In this paper, we propose a general-purpose adaptive antithetic sampling framework. We provide gradient-based and gradient-free methods to train the samplers such that they reduce variance while ensuring that the underlying Monte Carlo estimator is provably unbiased. We demonstrate the effectiveness of our approach on Bayesian inference and generative model training, where it reduces variance and improves task performance with little computational overhead.
****
Adversarial Online Learning with noise
Alon Resler, Yishay Mansour
We present and study models of adversarial online learning where the feedback observed by the learner is noisy, and the feedback is either full information feedback or bandit feedback. Specifically, we consider binary losses xored with the noise, which is a Bernoulli random variable. We consider both a constant noise rate and a variable noise rate. Our main results are tight regret bounds for learning with noise in the adversarial online learning model.
****
A Polynomial Time MCMC Method for Sampling from Continuous Determinantal Point Processes
Alireza Rezaei, Shayan Oveis Gharan
We study the Gibbs sampling algorithm for discrete and continuous k-determinantal point processes. We show that in both cases, the spectral gap of the chain is bounded by a polynomial of k and it is independent of the size of the domain. As an immediate corollary, we obtain sublinear time algorithms for sampling from discrete k-DPPs given access to polynomially many processors. In the continuous setting, our result leads to the first class of rigorously analyzed efficient algorithms to generate random samples of continuous k-DPPs. We achieve this by showing that the Gibbs sampler for a large family of continuous k-DPPs can be simulated efficiently when the spectrum is not concentrated on the top k eigenvalues.
****
A Persistent Weisfeiler-Lehman Procedure for Graph Classification
Bastian Rieck, Christian Bock, Karsten Borgwardt
The Weisfeiler–Lehman graph kernel exhibits competitive performance in many graph classification tasks. However, its subtree features are not able to capture connected components and cycles, topological features known for characterising graphs. To extract such features, we leverage propagated node label information and transform unweighted graphs into metric ones. This permits us to augment the subtree features with topological information obtained using persistent homology, a concept from topological data analysis. Our method, which we formalise as a generalisation of Weisfeiler–Lehman subtree features, exhibits favourable classification accuracy and its improvements in predictive performance are mainly driven by including cycle information.
****
Efficient learning of smooth probability functions from Bernoulli tests with guarantees
Paul Rolland, Ali Kavis, Alexander Immer, Adish Singla, Volkan Cevher
We study the fundamental problem of learning an unknown, smooth probability function via point-wise Bernoulli tests. We provide a scalable algorithm for efficiently solving this problem with rigorous guarantees. In particular, we prove the convergence rate of our posterior update rule to the true probability function in L2-norm. Moreover, we allow the Bernoulli tests to depend on contextual features, and provide a modified inference engine with provable guarantees for this novel setting. Numerical results show that the empirical convergence rates match the theory, and illustrate the superiority of our approach in handling contextual features over the state-of-the-art.
****

Separating value functions across time-scales
Joshua Romoff, Peter Henderson, Ahmed Touati, Emma Brunskill, Joelle Pineau, Yan
n Ollivier
In many finite horizon episodic reinforcement learning (RL) settings, it is desi
rable to optimize for the undiscounted return - in settings like Atari, for inst
ance, the goal is to collect the most points while staying alive in the long run
. Yet, it may be difficult (or even intractable) mathematically to learn with th
is target. As such, temporal discounting is often applied to optimize over a sho
rter effective planning horizon. This comes at the cost of potentially biasing t
he optimization target away from the undiscounted goal. In settings where this b
ias is unacceptable - where the system must optimize for longer horizons at high
er discounts - the target of the value function approximator may increase in var
iance leading to difficulties in learning. We present an extension of temporal d
ifference (TD) learning, which we call TD(Δ), that breaks down a value fu
nction into a series of components based on the differences between value functi
ons with smaller discount factors. The separation of a longer horizon value func
tion into these components has useful properties in scalability and performance.
 We discuss these properties and show theoretic and empirical improvements over
standard TD learning in certain settings.
****
Online Convex Optimization in Adversarial Markov Decision Processes
Aviv Rosenberg, Yishay Mansour
We consider online learning in episodic loop-free Markov decision processes (MDP
s), where the loss function can change arbitrarily between episodes, and the tra
nsition function is not known to the learner. We show $\tilde{O}(L|X|\sqrt{|A|T}
)$ regret bound, where T is the number of episodes, X is the state space, $A
$ is the action space, and L is the length of each episode. Our online algorit
hm is implemented using entropic regularization methodology, which allows to ext
end the original adversarial MDP model to handle convex performance criteria (di
fferent ways to aggregate the losses of a single episode) , as well as improve p
revious regret bounds.
****
Good Initializations of Variational Bayes for Deep Models
Simone Rossi, Pietro Michiardi, Maurizio Filippone
Stochastic variational inference is an established way to carry out approximate
Bayesian inference for deep models flexibly and at scale. While there have been
effective proposals for good initializations for loss minimization in deep learn
ing, far less attention has been devoted to the issue of initialization of stoch
astic variational inference. We address this by proposing a novel layer-wise ini
tialization strategy based on Bayesian linear models. The proposed method is ext
ensively validated on regression and classification tasks, including Bayesian De
ep Nets and Conv Nets, showing faster and better convergence compared to alterna
tives inspired by the literature on initializations for loss minimization.
****
The Odds are Odd: A Statistical Test for Detecting Adversarial Examples
Kevin Roth, Yannic Kilcher, Thomas Hofmann
We investigate conditions under which test statistics exist that can reliably de
tect examples, which have been adversarially manipulated in a white-box attack.
These statistics can be easily computed and calibrated by randomly corrupting in
puts. They exploit certain anomalies that adversarial attacks introduce, in part
icular if they follow the paradigm of choosing perturbations optimally under p-n
orm constraints. Access to the log-odds is the only requirement to defend models
. We justify our approach empirically, but also provide conditions under which d
etectability via the suggested test statistics is guaranteed to be effective. In
 our experiments, we show that it is even possible to correct test time predicti
ons for adversarial attacks with high accuracy.
****
Neuron birth-death dynamics accelerates gradient descent and converges asymptoti
cally
Grant Rotskoff, Samy Jelassi, Joan Bruna, Eric Vanden-Eijnden

Neural networks with a large number of parameters admit a mean-field description, which has recently served as a theoretical explanation for the favorable training properties of models with a large number of parameters. In this regime, gradient descent obeys a deterministic partial differential equation (PDE) that converges to a globally optimal solution for networks with a single hidden layer under appropriate assumptions. In this work, we propose a non-local mass transport dynamics that leads to a modified PDE with the same minimizer. We implement this non-local dynamics as a stochastic neuronal birth/death process and we prove that it accelerates the rate of convergence in the mean-field limit. We subsequently realize this PDE with two classes of numerical schemes that converge to the mean-field equation, each of which can easily be implemented for neural networks with finite numbers of parameters. We illustrate our algorithms with two models to provide intuition for the mechanism through which convergence is accelerated.
****

Iterative Linearized Control: Stable Algorithms and Complexity Guarantees

Vincent Roulet, Siddhartha Srinivasa, Dmitriy Drusvyatskiy, Zaid Harchaoui

We examine popular gradient-based algorithms for nonlinear control in the light of the modern complexity analysis of first-order optimization algorithms. The examination reveals that the complexity bounds can be clearly stated in terms of calls to a computational oracle related to dynamic programming and implementable by gradient back-propagation using machine learning software libraries such as PyTorch or TensorFlow. Finally, we propose a regularized Gauss-Newton algorithm enjoying worst-case complexity bounds and improved convergence behavior in practice. The software library based on PyTorch is publicly available.
****

Statistics and Samples in Distributional Reinforcement Learning

Mark Rowland, Robert Dadashi, Saurabh Kumar, Remi Munos, Marc G. Bellemare, Will Dabney

We present a unifying framework for designing and analysing distributional reinforcement learning (DRL) algorithms in terms of recursively estimating statistics of the return distribution. Our key insight is that DRL algorithms can be decomposed as the combination of some statistical estimator and a method for imputing a return distribution consistent with that set of statistics. With this new understanding, we are able to provide improved analyses of existing DRL algorithms as well as construct a new algorithm (EDRL) based upon estimation of the expectiles of the return distribution. We compare EDRL with existing methods on a variety of MDPs to illustrate concrete aspects of our analysis, and develop a deep RL variant of the algorithm, ER-DQN, which we evaluate on the Atari-57 suite of games.
****

A Contrastive Divergence for Combining Variational Inference and MCMC

Francisco Ruiz, Michalis Titsias

We develop a method to combine Markov chain Monte Carlo (MCMC) and variational inference (VI), leveraging the advantages of both inference approaches. Specifically, we improve the variational distribution by running a few MCMC steps. To make inference tractable, we introduce the variational contrastive divergence (VCD), a new divergence that replaces the standard Kullback-Leibler (KL) divergence used in VI. The VCD captures a notion of discrepancy between the initial variational distribution and its improved version (obtained after running the MCMC steps), and it converges asymptotically to the symmetrized KL divergence between the variational distribution and the posterior of interest. The VCD objective can be optimized efficiently with respect to the variational parameters via stochastic optimization. We show experimentally that optimizing the VCD leads to better predictive performance on two latent variable models: logistic matrix factorization and variational autoencoders (VAEs).
****

Plug-and-Play Methods Provably Converge with Properly Trained Denoisers

Ernest Ryu, Jialin Liu, Sicheng Wang, Xiaohan Chen, Zhangyang Wang, Wotao Yin

Plug-and-play (PnP) is a non-convex framework that integrates modern denoising priors, such as BM3D or deep learning-based denoisers, into ADMM or other proxima

l algorithms. An advantage of PnP is that one can use pre-trained denoisers when there is not sufficient data for end-to-end training. Although PnP has been recently studied extensively with great empirical success, theoretical analysis addressing even the most basic question of convergence has been insufficient. In this paper, we theoretically establish convergence of PnP-FBS and PnP-ADMM, without using diminishing stepsizes, under a certain Lipschitz condition on the denoisers. We then propose real spectral normalization, a technique for training deep learning-based denoisers to satisfy the proposed Lipschitz condition. Finally, we present experimental results validating the theory.
****
White-box vs Black-box: Bayes Optimal Strategies for Membership Inference
Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, Herve Jegou
Membership inference determines, given a sample and trained parameters of a machine learning model, whether the sample was part of the training set. In this paper, we derive the optimal strategy for membership inference with a few assumptions on the distribution of the parameters. We show that optimal attacks only depend on the loss function, and thus black-box attacks are as good as white-box attacks. As the optimal strategy is not tractable, we provide approximations of it leading to several inference methods, and show that existing membership inference methods are coarser approximations of this optimal strategy. Our membership attacks outperform the state of the art in various settings, ranging from a simple logistic regression to more complex architectures and datasets, such as ResNet-101 and Imagenet.
****
An Optimal Private Stochastic-MAB Algorithm based on Optimal Private Stopping Rule
Touqir Sajed, Or Sheffet
We present a provably optimal differentially private algorithm for the stochastic multi-arm bandit problem, as opposed to the private analogue of the UCB-algorithm (Mishra and Thakurta, 2015; Tossou and Dimitrakakis, 2016) which doesn’t meet the recently discovered lower-bound of $\Omega \left(\frac{K\log(T)}{\epsilon} \right)$ (Shariff and Sheffet, 2018). Our construction is based on a different algorithm, Successive Elimination (Even-Dar et al., 2002), that repeatedly pulls all remaining arms until an arm is found to be suboptimal and is then eliminated. In order to devise a private analogue of Successive Elimination we visit the problem of private stopping rule, that takes as input a stream of i.i.d samples from an unknown distribution and returns a multiplicative $(1 \pm \alpha)$-approximation of the distribution’s mean, and prove the optimality of our private stopping rule. We then present the private Successive Elimination algorithm which meets both the non-private lower bound (Lai and Robbins, 1985) and the above-mentioned private lower bound. We also compare empirically the performance of our algorithm with the private UCB algorithm.
****
Deep Gaussian Processes with Importance-Weighted Variational Inference
Hugh Salimbeni, Vincent Dutordoir, James Hensman, Marc Deisenroth
Deep Gaussian processes (DGPs) can model complex marginal densities as well as complex mappings. Non-Gaussian marginals are essential for modelling real-world data, and can be generated from the DGP by incorporating uncorrelated variables to the model. Previous work in the DGP model has introduced noise additively, and used variational inference with a combination of sparse Gaussian processes and mean-field Gaussians for the approximate posterior. Additive noise attenuates the signal, and the Gaussian form of variational distribution may lead to an inaccurate posterior. We instead incorporate noisy variables as latent covariates, and propose a novel importance-weighted objective, which leverages analytic results and provides a mechanism to trade off computation for improved accuracy. Our results demonstrate that the importance-weighted objective works well in practice and consistently outperforms classical variational inference, especially for deeper models.
****

Multivariate Submodular Optimization
Richard Santiago, F. Bruce Shepherd
Submodular functions have found a wealth of new applications in data science and machine learning models in recent years. This has been coupled with many algorithmic advances in the area of submodular optimization: (SO) $\min/\max f(S): S \in \mathcal{F}$, where $\mathcal{F}$ is a given family of feasible sets over a ground set V and $f:2^V \rightarrow \mathbb{R}$ is submodular. In this work we focus on a more general class of multivariate submodular optimization (MVSO) problems: $\min/\max f (S_1,S_2,\ldots,S_k): S_1 \uplus S_2 \uplus \cdots \uplus S_k \in \mathcal{F}$. Here we use $\uplus$ to denote union of disjoint sets and hence this model is attractive where resources are being allocated across k agents, who share a “joint” multivariate nonnegative objective $f(S_1,S_2,\ldots,S_k)$ that captures some type of submodularity (i.e. diminishing returns) property. We provide some explicit examples and potential applications for this new framework. For maximization, we show that practical algorithms such as accelerated greedy variants and distributed algorithms achieve good approximation guarantees for very general families (such as matroids and p-systems). For arbitrary families, we show that monotone (resp. nonmonotone) MVSO admits an $\alpha (1-1/e)$ (resp. $\alpha \cdot 0.385$) approximation whenever monotone (resp. nonmonotone) SO admits an α-approximation over the multilinear formulation. This substantially expands the family of tractable models. On the minimization side we give essentially optimal approximations in terms of the curvature of f.
****

Near optimal finite time identification of arbitrary linear dynamical systems
Tuhin Sarkar, Alexander Rakhlin
We derive finite time error bounds for estimating general linear time-invariant (LTI) systems from a single observed trajectory using the method of least squares. We provide the first analysis of the general case when eigenvalues of the LTI system are arbitrarily distributed in three regimes: stable, marginally stable, and explosive. Our analysis yields sharp upper bounds for each of these cases separately. We observe that although the underlying process behaves quite differently in each of these three regimes, the systematic analysis of a self–normalized martingale difference term helps bound identification error up to logarithmic factors of the lower bound. On the other hand, we demonstrate that the least squares solution may be statistically inconsistent under certain conditions even when the signal-to-noise ratio is high.
****

Breaking Inter-Layer Co-Adaptation by Classifier Anonymization
Ikuro Sato, Kohta Ishikawa, Guoqing Liu, Masayuki Tanaka
This study addresses an issue of co-adaptation between a feature extractor and a classifier in a neural network. A naive joint optimization of a feature extractor and a classifier often brings situations in which an excessively complex feature distribution adapted to a very specific classifier degrades the test performance. We introduce a method called Feature-extractor Optimization through Classifier Anonymization (FOCA), which is designed to avoid an explicit co-adaptation between a feature extractor and a particular classifier by using many randomly-generated, weak classifiers during optimization. We put forth a mathematical proposition that states the FOCA features form a point-like distribution within the same class in a class-separable fashion under special conditions. Real-data experiments under more general conditions provide supportive evidences.
****

A Theoretical Analysis of Contrastive Unsupervised Representation Learning
Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, Hrishikesh Khandeparkar
Recent empirical works have successfully used unlabeled data to learn feature representations that are broadly useful in downstream classification tasks. Several of these methods are reminiscent of the well-known word2vec embedding algorithm: leveraging availability of pairs of semantically “similar" data points and “negative samples," the learner forces the inner product of representations of similar pairs with each other to be higher on average than with negative samples. T

he current paper uses the term contrastive learning for such algorithms and presents a theoretical framework for analyzing them by introducing latent classes and hypothesizing that semantically similar points are sampled from the same latent class. This framework allows us to show provable guarantees on the performance of the learned representations on the average classification task that is comprised of a subset of the same set of latent classes. Our generalization bound also shows that learned representations can reduce (labeled) sample complexity on downstream tasks. We conduct controlled experiments in both the text and image domains to support the theory.
****
Locally Private Bayesian Inference for Count Models
Aaron Schein, Zhiwei Steven Wu, Alexandra Schofield, Mingyuan Zhou, Hanna Wallach
We present a general and modular method for privacy-preserving Bayesian inference for Poisson factorization, a broad class of models that includes some of the most widely used models in the social sciences. Our method satisfies limited-precision local privacy, a generalization of local differential privacy that we introduce to formulate appropriate privacy guarantees for sparse count data. We present an MCMC algorithm that approximates the posterior distribution over the latent variables conditioned on data that has been locally privatized by the geometric mechanism. Our method is based on two insights: 1) a novel reinterpretation of the geometric mechanism in terms of the Skellam distribution and 2) a general theorem that relates the Skellam and Bessel distributions. We demonstrate our method’s utility using two case studies that involve real-world email data. We show that our method consistently outperforms the commonly used naive approach, wherein inference proceeds as usual, treating the locally privatized data as if it were not privatized.
****
Weakly-Supervised Temporal Localization via Occurrence Count Learning
Julien Schroeter, Kirill Sidorov, David Marshall
We propose a novel model for temporal detection and localization which allows the training of deep neural networks using only counts of event occurrences as training labels. This powerful weakly-supervised framework alleviates the burden of the imprecise and time consuming process of annotating event locations in temporal data. Unlike existing methods, in which localization is explicitly achieved by design, our model learns localization implicitly as a byproduct of learning to count instances. This unique feature is a direct consequence of the model’s theoretical properties. We validate the effectiveness of our approach in a number of experiments (drum hit and piano onset detection in audio, digit detection in images) and demonstrate performance comparable to that of fully-supervised state-of-the-art methods, despite much weaker training requirements.
****
Discovering Context Effects from Raw Choice Data
Arjun Seshadri, Alex Peysakhovich, Johan Ugander
Many applications in preference learning assume that decisions come from the maximization of a stable utility function. Yet a large experimental literature shows that individual choices and judgements can be affected by “irrelevant” aspects of the context in which they are made. An important class of such contexts is the composition of the choice set. In this work, our goal is to discover such choice set effects from raw choice data. We introduce an extension of the Multinomial Logit (MNL) model, called the context dependent random utility model (CDM), which allows for a particular class of choice set effects. We show that the CDM can be thought of as a second-order approximation to a general choice system, can be inferred optimally using maximum likelihood and, importantly, is easily interpretable. We apply the CDM to both real and simulated choice data to perform principled exploratory analyses for the presence of choice set effects.
****
On the Feasibility of Learning, Rather than Assuming, Human Biases for Reward Inference
Rohin Shah, Noah Gundotra, Pieter Abbeel, Anca Dragan

Our goal is for agents to optimize the right reward function, despite how difficult it is for us to specify what that is. Inverse Reinforcement Learning (IRL) enables us to infer reward functions from demonstrations, but it usually assumes that the expert is noisily optimal. Real people, on the other hand, often have systematic biases: risk-aversion, myopia, etc. One option is to try to characterize these biases and account for them explicitly during learning. But in the era of deep learning, a natural suggestion researchers make is to avoid mathematical models of human behavior that are fraught with specific assumptions, and instead use a purely data-driven approach. We decided to put this to the test – rather than relying on assumptions about which specific bias the demonstrator has when planning, we instead learn the demonstrator's planning algorithm that they use to generate demonstrations, as a differentiable planner. Our exploration yielded mixed findings: on the one hand, learning the planner can lead to better reward inference than relying on the wrong assumption; on the other hand, this benefit is dwarfed by the loss we incur by going from an exact to a differentiable planner. This suggest that at least for the foreseeable future, agents need a middle ground between the flexibility of data-driven methods and the useful bias of known human biases. Code is available at https://tinyurl.com/learningbiases.
****
Exploration Conscious Reinforcement Learning Revisited
Lior Shani, Yonathan Efroni, Shie Mannor
The Exploration-Exploitation tradeoff arises in Reinforcement Learning when one cannot tell if a policy is optimal. Then, there is a constant need to explore new actions instead of exploiting past experience. In practice, it is common to resolve the tradeoff by using a fixed exploration mechanism, such as ϵ-greedy exploration or by adding Gaussian noise, while still trying to learn an optimal policy. In this work, we take a different approach and study exploration-conscious criteria, that result in optimal policies with respect to the exploration mechanism. Solving these criteria, as we establish, amounts to solving a surrogate Markov Decision Process. We continue and analyze properties of exploration-conscious optimal policies and characterize two general approaches to solve such criteria. Building on the approaches, we apply simple changes in existing tabular and deep Reinforcement Learning algorithms and empirically demonstrate superior performance relatively to their non-exploration-conscious counterparts, both for discrete and continuous action spaces.
****
Compressed Factorization: Fast and Accurate Low-Rank Factorization of Compressively-Sensed Data
Vatsal Sharan, Kai Sheng Tai, Peter Bailis, Gregory Valiant
What learning algorithms can be run directly on compressively-sensed data? In this work, we consider the question of accurately and efficiently computing low-rank matrix or tensor factorizations given data compressed via random projections. We examine the approach of first performing factorization in the compressed domain, and then reconstructing the original high-dimensional factors from the recovered (compressed) factors. In both the matrix and tensor settings, we establish conditions under which this natural approach will provably recover the original factors. While it is well-known that random projections preserve a number of geometric properties of a dataset, our work can be viewed as showing that they can also preserve certain solutions of non-convex, NP-Hard problems like non-negative matrix factorization. We support these theoretical results with experiments on synthetic data and demonstrate the practical applicability of compressed factorization on real-world gene expression and EEG time series datasets.
****
Conditional Independence in Testing Bayesian Networks
Yujia Shen, Haiying Huang, Arthur Choi, Adnan Darwiche
Testing Bayesian Networks (TBNs) were introduced recently to represent a set of distributions, one of which is selected based on the given evidence and used for reasoning. TBNs are more expressive than classical Bayesian Networks (BNs): Marginal queries correspond to multi-linear functions in BNs and to piecewise multi-linear functions in TBNs. Moreover, TBN queries are universal approximators, li

ke neural networks. In this paper, we study conditional independence in TBNs, showing that it can be inferred from d-separation as in BNs. We also study the role of TBN expressiveness and independence in dealing with the problem of learning with incomplete models (i.e., ones that miss nodes or edges from the data-generating model). Finally, we illustrate our results on a number of concrete examples, including a case study on Hidden Markov Models.
****
Learning to Clear the Market
Weiran Shen, Sebastien Lahaie, Renato Paes Leme
The problem of market clearing is to set a price for an item such that quantity demanded equals quantity supplied. In this work, we cast the problem of predicting clearing prices into a learning framework and use the resulting models to perform revenue optimization in auctions and markets with contextual information. The economic intuition behind market clearing allows us to obtain fine-grained control over the aggressiveness of the resulting pricing policy, grounded in theory. To evaluate our approach, we fit a model of clearing prices over a massive dataset of bids in display ad auctions from a major ad exchange. The learned prices outperform other modeling techniques in the literature in terms of revenue and efficiency trade-offs. Because of the convex nature of the clearing loss function, the convergence rate of our method is as fast as linear regression.
****
Mixture Models for Diverse Machine Translation: Tricks of the Trade
Tianxiao Shen, Myle Ott, Michael Auli, Marc’Aurelio Ranzato
Mixture models trained via EM are among the simplest, most widely used and well understood latent variable models in the machine learning literature. Surprisingly, these models have been hardly explored in text generation applications such as machine translation. In principle, they provide a latent variable to control generation and produce a diverse set of hypotheses. In practice, however, mixture models are prone to degeneracies—often only one component gets trained or the latent variable is simply ignored. We find that disabling dropout noise in responsibility computation is critical to successful training. In addition, the design choices of parameterization, prior distribution, hard versus soft EM and online versus offline assignment can dramatically affect model performance. We develop an evaluation protocol to assess both quality and diversity of generations against multiple references, and provide an extensive empirical study of several mixture model variants. Our analysis shows that certain types of mixture models are more robust and offer the best trade-off between translation quality and diversity compared to variational models and diverse decoding approaches.\footnote{Code to reproduce the results in this paper is available at \url{https://github.com/pytorch/fairseq}}
****
Hessian Aided Policy Gradient
Zebang Shen, Alejandro Ribeiro, Hamed Hassani, Hui Qian, Chao Mi
Reducing the variance of estimators for policy gradient has long been the focus of reinforcement learning research. ■While classic algorithms like REINFORCE find an ϵ-approximate first-order stationary point in $\OM({1}/{\epsilon^4})$ random trajectory simulations, no provable improvement on the complexity has been made so far. ■This paper presents a Hessian aided policy gradient method with the first improved sample complexity of $\OM({1}/{\epsilon^3})$. ■While our method exploits information from the policy Hessian, it can be implemented in linear time with respect to the parameter dimension and is hence applicable to sophisticated DNN parameterization. ■Simulations on standard tasks validate the efficiency of our method.
****
Learning with Bad Training Data via Iterative Trimmed Loss Minimization
Yanyao Shen, Sujay Sanghavi
In this paper, we study a simple and generic framework to tackle the problem of learning model parameters when a fraction of the training samples are corrupted. Our approach is motivated by a simple observation: in a variety of such settings, the evolution of training accuracy (as a function of training epochs) is diff

erent for clean samples and bad samples. We propose to iteratively minimize the trimmed loss, by alternating between (a) selecting samples with lowest current loss, and (b) retraining a model on only these samples. Analytically, we characterize the statistical performance and convergence rate of the algorithm for simple and natural linear and non-linear models. Experimentally, we demonstrate its effectiveness in three settings: (a) deep image classifiers with errors only in labels, (b) generative adversarial networks with bad training images, and (c) deep image classifiers with adversarial (image, label) pairs (i.e., backdoor attacks). For the well-studied setting of random label noise, our algorithm achieves state-of-the-art performance without having access to any a-priori guaranteed clean samples.
****

Replica Conditional Sequential Monte Carlo

Alex Shestopaloff, Arnaud Doucet

We propose a Markov chain Monte Carlo (MCMC) scheme to perform state inference in non-linear non-Gaussian state-space models. Current state-of-the-art methods to address this problem rely on particle MCMC techniques and its variants, such as the iterated conditional Sequential Monte Carlo (cSMC) scheme, which uses a Sequential Monte Carlo (SMC) type proposal within MCMC. A deficiency of standard SMC proposals is that they only use observations up to time t to propose states at time t when an entire observation sequence is available. More sophisticated SMC based on lookahead techniques could be used but they can be difficult to put in practice. We propose here replica cSMC where we build SMC proposals for one replica using information from the entire observation sequence by conditioning on the states of the other replicas. This approach is easily parallelizable and we demonstrate its excellent empirical performance when compared to the standard iterated cSMC scheme at fixed computational complexity.
****

Scalable Training of Inference Networks for Gaussian-Process Models

Jiaxin Shi, Mohammad Emtiyaz Khan, Jun Zhu

Inference in Gaussian process (GP) models is computationally challenging for large data, and often difficult to approximate with a small number of inducing points. We explore an alternative approximation that employs stochastic inference networks for a flexible inference. Unfortunately, for such networks, minibatch training is difficult to be able to learn meaningful correlations over function outputs for a large dataset. We propose an algorithm that enables such training by tracking a stochastic, functional mirror-descent algorithm. At each iteration, this only requires considering a finite number of input locations, resulting in a scalable and easy-to-implement algorithm. Empirical results show comparable and, sometimes, superior performance to existing sparse variational GP methods.
****

Fast Direct Search in an Optimally Compressed Continuous Target Space for Efficient Multi-Label Active Learning

Weishi Shi, Qi Yu

Active learning for multi-label classification poses fundamental challenges given the complex label correlations and a potentially large and sparse label space. We propose a novel CS-BPCA process that integrates compressed sensing and Bayesian principal component analysis to perform a two-level label transformation, resulting in an optimally compressed continuous target space. Besides leveraging correlation and sparsity of a large label space for effective compression, an optimal compressing rate and the relative importance of the resultant targets are automatically determined through Bayesian inference. Furthermore, the orthogonality of the transformed space completely decouples the correlations among targets, which significantly simplifies multi-label sampling in the target space. We define a novel sampling function that leverages a multi-output Gaussian Process (MOGP). Gradient-free optimization strategies are developed to achieve fast online hyper-parameter learning and model retraining for active learning. Experimental results over multiple real-world datasets and comparison with competitive multi-label active learning models demonstrate the effectiveness of the proposed framework.

****
Model-Based Active Exploration
Pranav Shyam, Wojciech Ja■kowski, Faustino Gomez
Efficient exploration is an unsolved problem in Reinforcement Learning which is usually addressed by reactively rewarding the agent for fortuitously encountering novel situations. This paper introduces an efficient active exploration algorithm, Model-Based Active eXploration (MAX), which uses an ensemble of forward models to plan to observe novel events. This is carried out by optimizing agent behaviour with respect to a measure of novelty derived from the Bayesian perspective of exploration, which is estimated using the disagreement between the futures predicted by the ensemble members. We show empirically that in semi-random discrete environments where directed exploration is critical to make progress, MAX is at least an order of magnitude more efficient than strong baselines. MAX scales to high-dimensional continuous environments where it builds task-agnostic models that can be used for any downstream task.
****
Rehashing Kernel Evaluation in High Dimensions
Paris Siminelakis, Kexin Rong, Peter Bailis, Moses Charikar, Philip Levis
Kernel methods are effective but do not scale well to large scale data, especially in high dimensions where the geometric data structures used to accelerate kernel evaluation suffer from the curse of dimensionality. Recent theoretical advances have proposed fast kernel evaluation algorithms leveraging hashing techniques with worst-case asymptotic improvements. However, these advances are largely confined to the theoretical realm due to concerns such as super-linear preprocessing time and diminishing gains in non-worst case datasets. In this paper, we close the gap between theory and practice by addressing these challenges via provable and practical procedures for adaptive sample size selection, preprocessing time reduction, and refined variance bounds that quantify the data-dependent performance of random sampling and hashing-based kernel evaluation methods. Our experiments show that these new tools offer up to $10\times$ improvement in evaluation time on a range of synthetic and real-world datasets.
****
Revisiting precision recall definition for generative modeling
Loic Simon, Ryan Webster, Julien Rabin
In this article we revisit the definition of Precision-Recall (PR) curves for generative models proposed by (Sajjadi et al., 2018). Rather than providing a scalar for generative quality, PR curves distinguish mode-collapse (poor recall) and bad quality (poor precision). We first generalize their formulation to arbitrary measures hence removing any restriction to finite support. We also expose a bridge between PR curves and type I and type II error (a.k.a. false detection and rejection) rates of likelihood ratio classifiers on the task of discriminating between samples of the two distributions. Building upon this new perspective, we propose a novel algorithm to approximate precision-recall curves, that shares some interesting methodological properties with the hypothesis testing technique from (Lopez-Paz & Oquab, 2017). We demonstrate the interest of the proposed formulation over the original approach on controlled multi-modal datasets.
****
First-Order Adversarial Vulnerability of Neural Networks and Input Dimension
Carl-Johann Simon-Gabriel, Yann Ollivier, Leon Bottou, Bernhard Schölkopf, David Lopez-Paz
Over the past few years, neural networks were proven vulnerable to adversarial images: targeted but imperceptible image perturbations lead to drastically different predictions. We show that adversarial vulnerability increases with the gradients of the training objective when viewed as a function of the inputs. Surprisingly, vulnerability does not depend on network topology: for many standard network architectures, we prove that at initialization, the L1-norm of these gradients grows as the square root of the input dimension, leaving the networks increasingly vulnerable with growing image size. We empirically show that this dimension-dependence persists after either usual or robust training, but gets attenuated with higher regularization.

****

Refined Complexity of PCA with Outliers

Kirill Simonov, Fedor Fomin, Petr Golovach, Fahad Panolan

Principal component analysis (PCA) is one of the most fundamental procedures in exploratory data analysis and is the basic step in applications ranging from quantitative finance and bioinformatics to image analysis and neuroscience. However, it is well-documented that the applicability of PCA in many real scenarios could be constrained by an "immune deficiency" to outliers such as corrupted observations. We consider the following algorithmic question about the PCA with outliers. For a set of n points in $\mathbb{R}^{d}$, how to learn a subset of points, say 1% of the total number of points, such that the remaining part of the points is best fit into some unknown r-dimensional subspace? We provide a rigorous algorithmic analysis of the problem. We show that the problem is solvable in time $n^{O(d^2)}$. In particular, for constant dimension the problem is solvable in polynomial time. We complement the algorithmic result by the lower bound, showing that unless Exponential Time Hypothesis fails, in time $f(d)n^{o(d)}$, for any function f of d, it is impossible not only to solve the problem exactly but even to approximate it within a constant factor.

****

A Tail-Index Analysis of Stochastic Gradient Noise in Deep Neural Networks

Umut Simsekli, Levent Sagun, Mert Gurbuzbalaban

The gradient noise (GN) in the stochastic gradient descent (SGD) algorithm is often considered to be Gaussian in the large data regime by assuming that the classical central limit theorem (CLT) kicks in. This assumption is often made for mathematical convenience, since it enables SGD to be analyzed as a stochastic differential equation (SDE) driven by a Brownian motion. We argue that the Gaussianity assumption might fail to hold in deep learning settings and hence render the Brownian motion-based analyses inappropriate. Inspired by non-Gaussian natural phenomena, we consider the GN in a more general context and invoke the generalized CLT (GCLT), which suggests that the GN converges to a heavy-tailed α-stable random variable. Accordingly, we propose to analyze SGD as an SDE driven by a Lévy motion. Such SDEs can incur ‘jumps’, which force the SDE transition from narrow minima to wider minima, as proven by existing metastability theory. To validate the α-stable assumption, we conduct experiments on common deep learning scenarios and show that in all settings, the GN is highly non-Gaussian and admits heavy-tails. We investigate the tail behavior in varying network architectures and sizes, loss functions, and datasets. Our results open up a different perspective and shed more light on the belief that SGD prefers wide minima.

****

Non-Parametric Priors For Generative Adversarial Networks

Rajhans Singh, Pavan Turaga, Suren Jayasuriya, Ravi Garg, Martin Braun

The advent of generative adversarial networks (GAN) has enabled new capabilities in synthesis, interpolation, and data augmentation heretofore considered very challenging. However, one of the common assumptions in most GAN architectures is the assumption of simple parametric latent-space distributions. While easy to implement, a simple latent-space distribution can be problematic for uses such as interpolation. This is due to distributional mismatches when samples are interpolated in the latent space. We present a straightforward formalization of this problem; using basic results from probability theory and off-the-shelf-optimization tools, we develop ways to arrive at appropriate non-parametric priors. The obtained prior exhibits unusual qualitative properties in terms of its shape, and quantitative benefits in terms of lower divergence with its mid-point distribution. We demonstrate that our designed prior helps improve image generation along any Euclidean straight line during interpolation, both qualitatively and quantitatively, without any additional training or architectural modifications. The proposed formulation is quite flexible, paving the way to impose newer constraints on the latent-space statistics.

****

Understanding Impacts of High-Order Loss Approximations and Features in Deep Learning Interpretation

Sahil Singla, Eric Wallace, Shi Feng, Soheil Feizi
Current saliency map interpretations for neural networks generally rely on two k
ey assumptions. First, they use first-order approximations of the loss function,
 neglecting higher-order terms such as the loss curvature. Second, they evaluate
 each feature’s importance in isolation, ignoring feature interdependencies. Thi
s work studies the effect of relaxing these two assumptions. First, we character
ize a closed-form formula for the input Hessian matrix of a deep ReLU network. U
sing this formula, we show that, for classification problems with many classes,
if a prediction has high probability then including the Hessian term has a small
 impact on the interpretation. We prove this result by demonstrating that these
conditions cause the Hessian matrix to be approximately rank one and its leading
 eigenvector to be almost parallel to the gradient of the loss. We empirically v
alidate this theory by interpreting ImageNet classifiers. Second, we incorporate
 feature interdependencies by calculating the importance of group-features using
 a sparsity regularization term. We use an L0 - L1 relaxation technique along wi
th proximal gradient descent to efficiently compute group-feature importance val
ues. Our empirical results show that our method significantly improves deep lear
ning interpretations.
****
kernelPSI: a Post-Selection Inference Framework for Nonlinear Variable Selection
Lotfi Slim, Clément Chatelain, Chloe-Agathe Azencott, Jean-Philippe Vert
Model selection is an essential task for many applications in scientific discove
ry. The most common approaches rely on univariate linear measures of association
 between each feature and the outcome. Such classical selection procedures fail
to take into account nonlinear effects and interactions between features. Kernel
-based selection procedures have been proposed as a solution. However, current s
trategies for kernel selection fail to measure the significance of a joint model
 constructed through the combination of the basis kernels. In the present work,
we exploit recent advances in post-selection inference to propose a valid statis
tical test for the association of a joint model of the selected kernels with the
 outcome. The kernels are selected via a step-wise procedure which we model as a
 succession of quadratic constraints in the outcome variable.
****
GEOMetrics: Exploiting Geometric Structure for Graph-Encoded Objects
Edward Smith, Scott Fujimoto, Adriana Romero, David Meger
Mesh models are a promising approach for encoding the structure of 3D objects. C
urrent mesh reconstruction systems predict uniformly distributed vertex location
s of a predetermined graph through a series of graph convolutions, leading to co
mpromises with respect to performance or resolution. In this paper, we argue tha
t the graph representation of geometric objects allows for additional structure,
 which should be leveraged for enhanced reconstruction. Thus, we propose a syste
m which properly benefits from the advantages of the geometric structure of grap
h-encoded objects by introducing (1) a graph convolutional update preserving ver
tex information; (2) an adaptive splitting heuristic allowing detail to emerge;
and (3) a training objective operating both on the local surfaces defined by ver
tices as well as the global structure defined by the mesh. Our proposed method i
s evaluated on the task of 3D object reconstruction from images with the ShapeNe
t dataset, where we demonstrate state of the art performance, both visually and
numerically, while having far smaller space requirements by generating adaptive
meshes.
****
The Evolved Transformer
David So, Quoc Le, Chen Liang
Recent works have highlighted the strength of the Transformer architecture on se
quence tasks while, at the same time, neural architecture search (NAS) has begun
 to outperform human-designed models. Our goal is to apply NAS to search for a b
etter alternative to the Transformer. We first construct a large search space in
spired by the recent advances in feed-forward sequence models and then run evolu
tionary architecture search with warm starting by seeding our initial population
 with the Transformer. To directly search on the computationally expensive WMT 2

014 English-German translation task, we develop the Progressive Dynamic Hurdles method, which allows us to dynamically allocate more resources to more promising candidate models. The architecture found in our experiments – the Evolved Transformer – demonstrates consistent improvement over the Transformer on four well-established language tasks: WMT 2014 English-German, WMT 2014 English-French, WMT 2014 English-Czech and LM1B. At a big model size, the Evolved Transformer establishes a new state-of-the-art BLEU score of 29.8 on WMT'14 English-German; at smaller sizes, it achieves the same quality as the original "big" Transformer with 37.6% less parameters and outperforms the Transformer by 0.7 BLEU at a mobile-friendly model size of 7M parameters.
****
QTRAN: Learning to Factorize with Transformation for Cooperative Multi-Agent Reinforcement Learning
Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, Yung Yi
We explore value-based solutions for multi-agent reinforcement learning (MARL) tasks in the centralized training with decentralized execution (CTDE) regime popularized recently. However, VDN and QMIX are representative examples that use the idea of factorization of the joint action-value function into individual ones for decentralized execution. VDN and QMIX address only a fraction of factorizable MARL tasks due to their structural constraint in factorization such as additivity and monotonicity. In this paper, we propose a new factorization method for MARL, QTRAN, which is free from such structural constraints and takes on a new approach to transforming the original joint action-value function into an easily factorizable one, with the same optimal actions. QTRAN guarantees more general factorization than VDN or QMIX, thus covering a much wider class of MARL tasks than does previous methods. Our experiments for the tasks of multi-domain Gaussian-squeeze and modified predator-prey demonstrate QTRAN's superior performance with especially larger margins in games whose payoffs penalize non-cooperative behavior more aggressively.
****
Distribution calibration for regression
Hao Song, Tom Diethe, Meelis Kull, Peter Flach
We are concerned with obtaining well-calibrated output distributions from regression models. Such distributions allow us to quantify the uncertainty that the model has regarding the predicted target value. We introduce the novel concept of distribution calibration, and demonstrate its advantages over the existing definition of quantile calibration. We further propose a post-hoc approach to improving the predictions from previously trained regression models, using multi-output Gaussian Processes with a novel Beta link function. The proposed method is experimentally verified on a set of common regression models and shows improvements for both distribution-level and quantile-level calibration.
****
SELFIE: Refurbishing Unclean Samples for Robust Deep Learning
Hwanjun Song, Minseok Kim, Jae-Gil Lee
Owing to the extremely high expressive power of deep neural networks, their side effect is to totally memorize training data even when the labels are extremely noisy. To overcome overfitting on the noisy labels, we propose a novel robust training method called SELFIE. Our key idea is to selectively refurbish and exploit unclean samples that can be corrected with high precision, thereby gradually increasing the number of available training samples. Taking advantage of this design, SELFIE effectively prevents the risk of noise accumulation from the false correction and fully exploits the training data. To validate the superiority of SELFIE, we conducted extensive experimentation using four real-world or synthetic data sets. The result showed that SELFIE remarkably improved absolute test error compared with two state-of-the-art methods.
****
Revisiting the Softmax Bellman Operator: New Benefits and New Perspective
Zhao Song, Ron Parr, Lawrence Carin
The impact of softmax on the value function itself in reinforcement learning (RL) is often viewed as problematic because it leads to sub-optimal value (or Q) fu

nctions and interferes with the contraction properties of the Bellman operator.
Surprisingly, despite these concerns, and independent of its effect on explorati
on, the softmax Bellman operator when combined with Deep Q-learning, leads to Q-
functions with superior policies in practice, even outperforming its double Q-le
arning counterpart. To better understand how and why this occurs, we revisit the
oretical properties of the softmax Bellman operator, and prove that (i) it conve
rges to the standard Bellman operator exponentially fast in the inverse temperat
ure parameter, and (ii) the distance of its Q function from the optimal one can
be bounded. These alone do not explain its superior performance, so we also show
 that the softmax operator can reduce the overestimation error, which may give s
ome insight into why a sub-optimal operator leads to better performance in the p
resence of value function approximation. A comparison among different Bellman op
erators is then presented, showing the trade-offs when selecting them.
****
MASS: Masked Sequence to Sequence Pre-training for Language Generation
Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu
Pre-training and fine-tuning, e.g., BERT \citep{devlin2018bert}, have achieved g
reat success in language understanding by transferring knowledge from rich-resou
rce pre-training task to the low/zero-resource downstream tasks. Inspired by the
 success of BERT, we propose MAsked Sequence to Sequence pre-training (MASS) for
 the encoder-decoder based language generation tasks. MASS adopts the encoder-de
coder framework to reconstruct a sentence fragment given the remaining part of t
he sentence: its encoder takes a sentence with randomly masked fragment (several
 consecutive tokens) as input, and its decoder tries to predict this masked frag
ment. In this way, MASS can jointly train the encoder and decoder to develop the
 capability of representation extraction and language modeling. By further fine-
tuning on a variety of zero/low-resource language generation tasks, including ne
ural machine translation, text summarization and conversational response generat
ion (3 tasks and totally 8 datasets), MASS achieves significant improvements ove
r the baselines without pre-training or with other pre-training methods. Especia
lly, we achieve the state-of-the-art accuracy (30.02 in terms of BLEU score) on
the unsupervised English-French translation, even beating the early attention-ba
sed supervised model \citep{bahdanau2015neural}.
****
Dual Entangled Polynomial Code: Three-Dimensional Coding for Distributed Matrix
Multiplication
Pedro Soto, Jun Li, Xiaodi Fan
Matrix multiplication is a fundamental building block in various machine learnin
g algorithms. When the matrix comes from a large dataset, the multiplication can
 be split into multiple tasks which calculate the multiplication of submatrices
on different nodes. As some nodes may be stragglers, coding schemes have been pr
oposed to tolerate stragglers in such distributed matrix multiplication. However
, existing coding schemes typically split the matrices in only one or two dimens
ions, limiting their capabilities to handle large-scale matrix multiplication. T
hree-dimensional coding, however, does not have any code construction that achie
ves the optimal number of tasks required for decoding, with the best result achi
eved by entangled polynomial (EP) codes. In this paper, we propose dual entangle
d polynomial (DEP) codes that require around 25% fewer tasks than EP codes by ex
ecuting two matrix multiplications on each task. With experiments in a real clou
d environment, we show that DEP codes can also save the decoding overhead and me
mory consumption of tasks.
****
Compressing Gradient Optimizers via Count-Sketches
Ryan Spring, Anastasios Kyrillidis, Vijai Mohan, Anshumali Shrivastava
Many popular first-order optimization methods accelerate the convergence rate of
 deep learning models. However, these algorithms require auxiliary variables, wh
ich cost additional memory proportional to the number of parameters in the model
. The problem is becoming more severe as models grow larger to learn from comple
x, large-scale datasets. Our proposed solution is to maintain a linear sketch to
 compress the auxiliary variables. Our approach has the same performance as the

full-sized baseline, while using less space for the auxiliary variables. Theoret
ically, we prove that count-sketch optimization maintains the SGD convergence ra
te, while gracefully reducing memory usage for large-models. We show a rigorous
evaluation on popular architectures such as ResNet-18 and Transformer-XL. On the
1-Billion Word dataset, we save 25% of the memory used during training (7.7 GB
instead of 10.8 GB) with minimal accuracy and performance loss. For an Amazon ex
treme classification task with over 49.5 million classes, we also reduce the tra
ining time by 38%, by increasing the mini-batch size 3.5x using our count-sketch
optimizer.
****
Escaping Saddle Points with Adaptive Gradient Methods
Matthew Staib, Sashank Reddi, Satyen Kale, Sanjiv Kumar, Suvrit Sra
Adaptive methods such as Adam and RMSProp are widely used in deep learning but a
re not well understood. In this paper, we seek a crisp, clean and precise charac
terization of their behavior in nonconvex settings. To this end, we first provid
e a novel view of adaptive methods as preconditioned SGD, where the precondition
er is estimated in an online manner. By studying the preconditioner on its own,
we elucidate its purpose: it rescales the stochastic gradient noise to be isotro
pic near stationary points, which helps escape saddle points. Furthermore, we sh
ow that adaptive methods can efficiently estimate the aforementioned preconditio
ner. By gluing together these two components, we provide the first (to our knowl
edge) second-order convergence result for any adaptive method. The key insight f
rom our analysis is that, compared to SGD, adaptive methods escape saddle points
faster, and can converge faster overall to second-order stationary points.
****
Faster Attend-Infer-Repeat with Tractable Probabilistic Models
Karl Stelzner, Robert Peharz, Kristian Kersting
The recent Attend-Infer-Repeat (AIR) framework marks a milestone in structured p
robabilistic modeling, as it tackles the challenging problem of unsupervised sce
ne understanding via Bayesian inference. AIR expresses the composition of visual
scenes from individual objects, and uses variational autoencoders to model the
appearance of those objects. However, inference in the overall model is highly i
ntractable, which hampers its learning speed and makes it prone to suboptimal so
lutions. In this paper, we show that the speed and robustness of learning in AIR
can be considerably improved by replacing the intractable object representation
s with tractable probabilistic models. In particular, we opt for sum-product net
works (SPNs), expressive deep probabilistic models with a rich set of tractable
inference routines. The resulting model, called SuPAIR, learns an order of magni
tude faster than AIR, treats object occlusions in a consistent manner, and allow
s for the inclusion of a background noise model, improving the robustness of Bay
esian scene understanding.
****
Insertion Transformer: Flexible Sequence Generation via Insertion Operations
Mitchell Stern, William Chan, Jamie Kiros, Jakob Uszkoreit
We present the Insertion Transformer, an iterative, partially autoregressive mod
el for sequence generation based on insertion operations. Unlike typical autoreg
ressive models which rely on a fixed, often left-to-right ordering of the output
, our approach accommodates arbitrary orderings by allowing for tokens to be ins
erted anywhere in the sequence during decoding. This flexibility confers a numbe
r of advantages: for instance, not only can our model be trained to follow speci
fic orderings such as left-to-right generation or a binary tree traversal, but i
t can also be trained to maximize entropy over all valid insertions for robustne
ss. In addition, our model seamlessly accommodates both fully autoregressive gen
eration (one insertion at a time) and partially autoregressive generation (simul
taneous insertions at multiple locations). We validate our approach by analyzing
its performance on the WMT 2014 English-German machine translation task under v
arious settings for training and decoding. We find that the Insertion Transforme
r outperforms many prior non-autoregressive approaches to translation at compara
ble or better levels of parallelism, and successfully recovers the performance o
f the original Transformer while requiring only logarithmically many iterations

during decoding.
****
BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning
Asa Cooper Stickland, Iain Murray
Multi-task learning shares information between related tasks, sometimes reducing the number of parameters required. State-of-the-art results across multiple natural language understanding tasks in the GLUE benchmark have previously used transfer from a single large task: unsupervised pre-training with BERT, where a separate BERT model was fine-tuned for each task. We explore multi-task approaches that share a \hbox{single} BERT model with a small number of additional task-specific parameters. Using new adaptation modules, PALs or ‘projected attention layers’, we match the performance of separately fine-tuned models on the GLUE benchmark with $\approx$7 times fewer parameters, and obtain state-of-the-art results on the Recognizing Textual Entailment dataset.
****
Learning Optimal Linear Regularizers
Matthew Streeter
We present algorithms for efficiently learning regularizers that improve generalization. Our approach is based on the insight that regularizers can be viewed as upper bounds on the generalization gap, and that reducing the slack in the bound can improve performance on test data. For a broad class of regularizers, the hyperparameters that give the best upper bound can be computed using linear programming. Under certain Bayesian assumptions, solving the LP lets us "jump" to the optimal hyperparameters given very limited data. This suggests a natural algorithm for tuning regularization hyperparameters, which we show to be effective on both real and synthetic data.
****
CAB: Continuous Adaptive Blending for Policy Evaluation and Learning
Yi Su, Lequn Wang, Michele Santacatterina, Thorsten Joachims
The ability to perform offline A/B-testing and off-policy learning using logged contextual bandit feedback is highly desirable in a broad range of applications, including recommender systems, search engines, ad placement, and personalized health care. Both offline A/B-testing and off-policy learning require a counterfactual estimator that evaluates how some new policy would have performed, if it had been used instead of the logging policy. In this paper, we identify a family of counterfactual estimators which subsumes most such estimators proposed to date. Our analysis of this family identifies a new estimator - called Continuous Adaptive Blending (CAB) - which enjoys many advantageous theoretical and practical properties. In particular, it can be substantially less biased than clipped Inverse Propensity Score (IPS) weighting and the Direct Method, and it can have less variance than Doubly Robust and IPS estimators. In addition, it is sub-differentiable such that it can be used for learning, unlike the SWITCH estimator. Experimental results show that CAB provides excellent evaluation accuracy and outperforms other counterfactual estimators in terms of learning performance.
****
Learning Distance for Sequences by Learning a Ground Metric
Bing Su, Ying Wu
Learning distances that operate directly on multi-dimensional sequences is challenging because such distances are structural by nature and the vectors in sequences are not independent. Generally, distances for sequences heavily depend on the ground metric between the vectors in sequences. We propose to learn the distance for sequences through learning a ground Mahalanobis metric for the vectors in sequences. The learning samples are sequences of vectors for which how the ground metric between vectors induces the overall distance is given, and the objective is that the distance induced by the learned ground metric produces large values for sequences from different classes and small values for those from the same class. We formulate the metric as a parameter of the distance, bring closer each sequence to an associated virtual sequence w.r.t. the distance to reduce the number of constraints, and develop a general iterative solution for any ground-me

tric-based sequence distance. Experiments on several sequence datasets demonstrate the effectiveness and efficiency of our method.
****
Contextual Memory Trees
Wen Sun, Alina Beygelzimer, Hal Daumé Iii, John Langford, Paul Mineiro
We design and study a Contextual Memory Tree (CMT), a learning memory controller that inserts new memories into an experience store of unbounded size. It operates online and is designed to efficiently query for memories from that store, supporting logarithmic time insertion and retrieval operations. Hence CMT can be integrated into existing statistical learning algorithms as an augmented memory unit without substantially increasing training and inference computation. Furthermore CMT operates as a reduction to classification, allowing it to benefit from advances in representation or architecture. We demonstrate the efficacy of CMT by augmenting existing multi-class and multi-label classification algorithms with CMT and observe statistical improvement. We also test CMT learning on several image-captioning tasks to demonstrate that it performs computationally better than a simple nearest neighbors memory system while benefitting from reward learning.
****
Provably Efficient Imitation Learning from Observation Alone
Wen Sun, Anirudh Vemula, Byron Boots, Drew Bagnell
We study Imitation Learning (IL) from Observations alone (ILFO) in large-scale MDPs. While most IL algorithms rely on an expert to directly provide actions to the learner, in this setting the expert only supplies sequences of observations. We design a new model-free algorithm for ILFO, Forward Adversarial Imitation Learning (FAIL), which learns a sequence of time-dependent policies by minimizing an Integral Probability Metric between the observation distributions of the expert policy and the learner. FAIL provably learns a near-optimal policy with a number of samples that is polynomial in all relevant parameters but independent of the number of unique observations. The resulting theory extends the domain of provably sample efficient learning algorithms beyond existing results that typically only consider tabular RL settings or settings that require access to a near-optimal reset distribution. We also demonstrate the efficacy ofFAIL on multiple OpenAI Gym control tasks.
****
Active Learning for Decision-Making from Imbalanced Observational Data
Iiris Sundin, Peter Schulam, Eero Siivola, Aki Vehtari, Suchi Saria, Samuel Kaski
Machine learning can help personalized decision support by learning models to predict individual treatment effects (ITE). This work studies the reliability of prediction-based decision-making in a task of deciding which action a to take for a target unit after observing its covariates $\tilde{x}$ and predicted outcomes $\hat{p}(\tilde{y} \mid \tilde{x}, a)$. An example case is personalized medicine and the decision of which treatment to give to a patient. A common problem when learning these models from observational data is imbalance, that is, difference in treated/control covariate distributions, which is known to increase the upper bound of the expected ITE estimation error. We propose to assess the decision-making reliability by estimating the ITE model’s Type S error rate, which is the probability of the model inferring the sign of the treatment effect wrong. Furthermore, we use the estimated reliability as a criterion for active learning, in order to collect new (possibly expensive) observations, instead of making a forced choice based on unreliable predictions. We demonstrate the effectiveness of this decision-making aware active learning in two decision-making tasks: in simulated data with binary outcomes and in a medical dataset with synthetic and continuous treatment outcomes.
****
Robustly Disentangled Causal Mechanisms: Validating Deep Representations for Interventional Robustness
Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, Stefan Bauer
The ability to learn disentangled representations that split underlying sources

of variation in high dimensional, unstructured data is important for data efficient and robust use of neural networks. While various approaches aiming towards this goal have been proposed in recent times, a commonly accepted definition and validation procedure is missing. We provide a causal perspective on representation learning which covers disentanglement and domain shift robustness as special cases. Our causal framework allows us to introduce a new metric for the quantitative evaluation of deep latent variable models. We show how this metric can be estimated from labeled observational data and further provide an efficient estimation algorithm that scales linearly in the dataset size.
****

Hyperbolic Disk Embeddings for Directed Acyclic Graphs

Ryota Suzuki, Ryusuke Takahama, Shun Onoda

Obtaining continuous representations of structural data such as directed acyclic graphs (DAGs) has gained attention in machine learning and artificial intelligence. However, embedding complex DAGs in which both ancestors and descendants of nodes are exponentially increasing is difficult. Tackling in this problem, we develop Disk Embeddings, which is a framework for embedding DAGs into quasi-metric spaces. Existing state-of-the-art methods, Order Embeddings and Hyperbolic Entailment Cones, are instances of Disk Embedding in Euclidean space and spheres respectively. Furthermore, we propose a novel method Hyperbolic Disk Embeddings to handle exponential growth of relations. The results of our experiments show that our Disk Embedding models outperform existing methods especially in complex DAGs other than trees.
****

Accelerated Flow for Probability Distributions

Amirhossein Taghvaei, Prashant Mehta

This paper presents a methodology and numerical algorithms for constructing accelerated gradient flows on the space of probability distributions. In particular, we extend the recent variational formulation of accelerated methods in (Wibisono et al., 2016) from vector valued variables to probability distributions. The variational problem is modeled as a mean-field optimal control problem. A quantitative estimate on the asymptotic convergence rate is provided based on a Lyapunov function construction, when the objective functional is displacement convex. An important special case is considered where the objective functional is the relative entropy. For this case, two numerical approximations are presented to implement the Hamilton’s equations as a system of N interacting particles. The algorithm is numerically illustrated and compared with the MCMC and Hamiltonian MCMC algorithms.
****

Equivariant Transformer Networks

Kai Sheng Tai, Peter Bailis, Gregory Valiant

How can prior knowledge on the transformation invariances of a domain be incorporated into the architecture of a neural network? We propose Equivariant Transformers (ETs), a family of differentiable image-to-image mappings that improve the robustness of models towards pre-defined continuous transformation groups. Through the use of specially-derived canonical coordinate systems, ETs incorporate functions that are equivariant by construction with respect to these transformations. We show empirically that ETs can be flexibly composed to improve model robustness towards more complicated transformation groups in several parameters. On a real-world image classification task, ETs improve the sample efficiency of ResNet classifiers, achieving relative improvements in error rate of up to 15% in the limited data regime while increasing model parameter count by less than 1%.
****

Making Deep Q-learning methods robust to time discretization

Corentin Tallec, Léonard Blier, Yann Ollivier

Despite remarkable successes, Deep Reinforce- ment Learning (DRL) is not robust to hyperparam- eterization, implementation details, or small envi- ronment changes (Henderson et al. 2017, Zhang et al. 2018). Overcoming such sensitivity is key to making DRL applicable to real world problems. In this paper, we identify sensitivity to time dis- cretization in near continuous-time environments as a cri

tical factor; this covers, e.g., changing the number of frames per second, or th e action frequency of the controller. Empirically, we find that Q-learning-based approaches such as Deep Q- learning (Mnih et al., 2015) and Deep Determinis- ti c Policy Gradient (Lillicrap et al., 2015) collapse with small time steps. Forma lly, we prove that Q-learning does not exist in continuous time. We detail a pri ncipled way to build an off-policy RL algorithm that yields similar performances over a wide range of time discretizations, and confirm this robustness empirica lly.
****
EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks
Mingxing Tan, Quoc Le
Convolutional Neural Networks (ConvNets) are commonly developed at a fixed resou rce budget, and then scaled up for better accuracy if more resources are given. In this paper, we systematically study model scaling and identify that carefully balancing network depth, width, and resolution can lead to better performance. Based on this observation, we propose a new scaling method that uniformly scales all dimensions of depth/width/resolution using a simple yet highly effective co mpound coefficient. We demonstrate the effectiveness of this method on MobileNet s and ResNet. To go even further, we use neural architecture search to design a new baseline network and scale it up to obtain a family of models, called Effici entNets, which achieve much better accuracy and efficiency than previous ConvNet s. In particular, our EfficientNet-B7 achieves stateof-the-art 84.4% top-1 / 97. 1% top-5 accuracy on ImageNet, while being 8.4x smaller and 6.1x faster on infer ence than the best existing ConvNet (Huang et al., 2018). Our EfficientNets also transfer well and achieve state-of-the-art accuracy on CIFAR-100 (91.7%), Flowe r (98.8%), and 3 other transfer learning datasets, with an order of magnitude fe wer parameters.
****
Hierarchical Decompositional Mixtures of Variational Autoencoders
Ping Liang Tan, Robert Peharz
Variational autoencoders (VAEs) have received considerable attention, since they allow us to learn expressive neural density estimators effectively and efficien tly. However, learning and inference in VAEs is still problematic due to the sen sitive interplay between the generative model and the inference network. Since t hese problems become generally more severe in high dimensions, we propose a nove l hierarchical mixture model over low-dimensional VAE experts. Our model decompo ses the overall learning problem into many smaller problems, which are coordinat ed by the hierarchical mixture, represented by a sum-product network. In experim ents we show that our models outperform classical VAEs on almost all of our expe rimental benchmarks. Moreover, we show that our model is highly data efficient a nd degrades very gracefully in extremely low data regimes.ow data regimes.
****
Mallows ranking models: maximum likelihood estimate and regeneration
Wenpin Tang
This paper is concerned with various Mallows ranking models. We study the statis tical properties of the MLE of Mallows’ ϕ model. We also make connections o f various Mallows ranking models, encompassing recent progress in mathematics. M otivated by the infinite top-t ranking model, we propose an algorithm to selec t the model size t automatically. The key idea relies on the renewal property of such an infinite random permutation. Our algorithm shows good performance on several data sets.
****
Correlated Variational Auto-Encoders
Da Tang, Dawen Liang, Tony Jebara, Nicholas Ruozzi
Variational Auto-Encoders (VAEs) are capable of learning latent representations for high dimensional data. However, due to the i.i.d. assumption, VAEs only opti mize the singleton variational distributions and fail to account for the correla tions between data points, which might be crucial for learning latent representa tions from dataset where a priori we know correlations exist. We propose Correla ted Variational Auto-Encoders (CVAEs) that can take the correlation structure in

to consideration when learning latent representations with VAEs. CVAEs apply a p
rior based on the correlation structure. To address the intractability introduce
d by the correlated prior, we develop an approximation by average of a set of tr
actable lower bounds over all maximal acyclic subgraphs of the undirected correl
ation graph. Experimental results on matching and link prediction on public benc
hmark rating datasets and spectral clustering on a synthetic dataset show the ef
fectiveness of the proposed method over baseline algorithms.
****
The Variational Predictive Natural Gradient
Da Tang, Rajesh Ranganath
Variational inference transforms posterior inference into parametric optimizatio
n thereby enabling the use of latent variable models where otherwise impractical
. However, variational inference can be finicky when different variational param
eters control variables that are strongly correlated under the model. Traditiona
l natural gradients based on the variational approximation fail to correct for c
orrelations when the approximation is not the true posterior. To address this, w
e construct a new natural gradient called the Variational Predictive Natural Gra
dient (VPNG). Unlike traditional natural gradients for variational inference, th
is natural gradient accounts for the relationship between model parameters and v
ariational parameters. We demonstrate the insight with a simple example as well
as the empirical value on a classification task, a deep generative model of imag
es, and probabilistic matrix factorization for recommendation.
****
DoubleSqueeze: Parallel Stochastic Gradient Descent with Double-pass Error-Compe
nsated Compression
Hanlin Tang, Chen Yu, Xiangru Lian, Tong Zhang, Ji Liu
A standard approach in large scale machine learning is distributed stochastic gr
adient training, which requires the computation of aggregated stochastic gradien
ts over multiple nodes on a network. Communication is a major bottleneck in such
 applications, and in recent years, compressed stochastic gradient methods such
as QSGD (quantized SGD) and sparse SGD have been proposed to reduce communicatio
n. It was also shown that error compensation can be combined with compression to
 achieve better convergence in a scheme that each node compresses its local stoc
hastic gradient and broadcast the result to all other nodes over the network in
a single pass. However, such a single pass broadcast approach is not realistic i
n many practical implementations. For example, under the popular parameter-serve
r model for distributed learning, the worker nodes need to send the compressed l
ocal gradients to the parameter server, which performs the aggregation. The para
meter server has to compress the aggregated stochastic gradient again before sen
ding it back to the worker nodes. In this work, we provide a detailed analysis o
n this two-pass communication model, with error-compensated compression both on
the worker nodes and on the parameter server. We show that the error-compensated
 stochastic gradient algorithm admits three very nice properties: 1) it is compa
tible with an arbitrary compression technique; 2) it admits an improved converge
nce rate than the non error-compensated stochastic gradient method such as QSGD
and sparse SGD; 3) it admits linear speedup with respect to the number of worker
s. The empirical study is also conducted to validate our theoretical results.
****
Adaptive Neural Trees
Ryutaro Tanno, Kai Arulkumaran, Daniel Alexander, Antonio Criminisi, Aditya Nori
Deep neural networks and decision trees operate on largely separate paradigms; t
ypically, the former performs representation learning with pre-specified archite
ctures, while the latter is characterised by learning hierarchies over pre-speci
fied features with data-driven architectures. We unite the two via adaptive neur
al trees (ANTs), a model that incorporates representation learning into edges, r
outing functions and leaf nodes of a decision tree, along with a backpropagation
-based training algorithm that adaptively grows the architecture from primitive
modules (e.g., convolutional layers). We demonstrate that, whilst achieving comp
etitive performance on classification and regression datasets, ANTs benefit from
 (i) lightweight inference via conditional computation, (ii) hierarchical separa

tion of features useful to the predictive task e.g. learning meaningful class as sociations, such as separating natural vs. man-made objects, and (iii) a mechani sm to adapt the architecture to the size and complexity of the training dataset.
****

Variational Annealing of GANs: A Langevin Perspective
Chenyang Tao, Shuyang Dai, Liqun Chen, Ke Bai, Junya Chen, Chang Liu, Ruiyi Zhan g, Georgiy Bobashev, Lawrence Carin Duke
The generative adversarial network (GAN) has received considerable attention rec ently as a model for data synthesis, without an explicit specification of a like lihood function. There has been commensurate interest in leveraging likelihood e stimates to improve GAN training. To enrich the understanding of this fast-growi ng yet almost exclusively heuristic-driven subject, we elucidate the theoretical roots of some of the empirical attempts to stabilize and improve GAN training w ith the introduction of likelihoods. We highlight new insights from variational theory of diffusion processes to derive a likelihood-based regularizing scheme f or GAN training, and present a novel approach to train GANs with an unnormalized distribution instead of empirical samples. To substantiate our claims, we provi de experimental evidence on how our theoretically-inspired new algorithms improv e upon current practice.
****

Predicate Exchange: Inference with Declarative Knowledge
Zenna Tavares, Javier Burroni, Edgar Minasyan, Armando Solar-Lezama, Rajesh Rang anath
Programming languages allow us to express complex predicates, but existing infer ence methods are unable to condition probabilistic models on most of them. To su pport a broader class of predicates, we develop an inference procedure called pr edicate exchange, which softens predicates. A soft predicate quantifies the exte nt to which values of model variables are consistent with its hard counterpart. We substitute the likelihood term in the Bayesian posterior with a soft predicat e, and develop a variant of replica exchange MCMC to draw posterior samples. We implement predicate exchange as a language agnostic tool which performs a nonsta ndard execution of a probabilistic program. We demonstrate the approach on seque nce models of health and inverse rendering.
****

The Natural Language of Actions
Guy Tennenholtz, Shie Mannor
We introduce Act2Vec, a general framework for learning context-based action repr esentation for Reinforcement Learning. Representing actions in a vector space he lp reinforcement learning algorithms achieve better performance by grouping simi lar actions and utilizing relations between different actions. We show how prior knowledge of an environment can be extracted from demonstrations and injected i nto action vector representations that encode natural compatible behavior. We th en use these for augmenting state representations as well as improving function approximation of Q-values. We visualize and test action embeddings in three doma ins including a drawing task, a high dimensional navigation task, and the large action space domain of StarCraft II.
****

Kernel Normalized Cut: a Theoretical Revisit
Yoshikazu Terada, Michio Yamamoto
In this paper, we study the theoretical properties of clustering based on the ke rnel normalized cut. Our first contribution is to derive a nonasymptotic upper b ound on the expected distortion rate of the kernel normalized cut. From this res ult, we show that the solution of the kernel normalized cut converges to that of the population-level weighted k-means clustering on a certain reproducing kerne l Hilbert space (RKHS). Our second contribution is the discover of the interesti ng fact that the population-level weighted k-means clustering in the RKHS is equ ivalent to the population-level normalized cut. Combining these results, we can see that the kernel normalized cut converges to the population-level normalized cut. The criterion of the population-level normalized cut can be considered as a n indivisibility of the population distribution, and this criterion plays an imp

ortant role in the theoretical analysis of spectral clustering in Schiebinger et al. (2015). We believe that our results will provide deep insights into the behavior of both normalized cut and spectral clustering.
****
Action Robust Reinforcement Learning and Applications in Continuous Control
Chen Tessler, Yonathan Efroni, Shie Mannor
A policy is said to be robust if it maximizes the reward while considering a bad, or even adversarial, model. In this work we formalize two new criteria of robustness to action uncertainty. Specifically, we consider two scenarios in which the agent attempts to perform an action $\action$, and (i) with probability α, an alternative adversarial action $\bar \action$ is taken, or (ii) an adversary adds a perturbation to the selected action in the case of continuous action space. We show that our criteria are related to common forms of uncertainty in robotics domains, such as the occurrence of abrupt forces, and suggest algorithms in the tabular case. Building on the suggested algorithms, we generalize our approach to deep reinforcement learning (DRL) and provide extensive experiments in the various MuJoCo domains. Our experiments show that not only does our approach produce robust policies, but it also improves the performance in the absence of perturbations. This generalization indicates that action-robustness can be thought of as implicit regularization in RL problems.
****
Concentration Inequalities for Conditional Value at Risk
Philip Thomas, Erik Learned-Miller
In this paper we derive new concentration inequalities for the conditional value at risk (CVaR) of a random variable, and compare them to the previous state of the art (Brown, 2007). We show analytically that our lower bound is strictly tighter than Brown’s, and empirically that this difference is significant. While our upper bound may be looser than Brown’s in some cases, we show empirically that in most cases our bound is significantly tighter. After discussing when each upper bound is superior, we conclude with empirical results which suggest that both of our bounds will often be significantly tighter than Brown’s.
****
Combating Label Noise in Deep Learning using Abstention
Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff Bilmes, Gopinath Chennupati, Jamal Mohd-Yusof
We introduce a novel method to combat label noise when training deep neural networks for classification. We propose a loss function that permits abstention during training thereby allowing the DNN to abstain on confusing samples while continuing to learn and improve classification performance on the non-abstained samples. We show how such a deep abstaining classifier (DAC) can be used for robust learning in the presence of different types of label noise. In the case of structured or systematic label noise {–} where noisy training labels or confusing examples are correlated with underlying features of the data{–} training with abstention enables representation learning for features that are associated with unreliable labels. In the case of unstructured (arbitrary) label noise, abstention during training enables the DAC to be used as an effective data cleaner by identifying samples that are likely to have label noise. We provide analytical results on the loss function behavior that enable dynamic adaption of abstention rates based on learning progress during training. We demonstrate the utility of the deep abstaining classifier for various image classification tasks under different types of label noise; in the case of arbitrary label noise, we show significant im- provements over previously published results on multiple image benchmarks.
****
ELF OpenGo: an analysis and open reimplementation of AlphaZero
Yuandong Tian, Jerry Ma, Qucheng Gong, Shubho Sengupta, Zhuoyuan Chen, James Pinkerton, Larry Zitnick
The AlphaGo, AlphaGo Zero, and AlphaZero series of algorithms are remarkable demonstrations of deep reinforcement learning’s capabilities, achieving superhuman performance in the complex game of Go with progressively increasing autonomy. However, many obstacles remain in the understanding of and usability of these prom

ising approaches by the research community. Toward elucidating unresolved myster
ies and facilitating future research, we propose ELF OpenGo, an open-source reim
plementation of the AlphaZero algorithm. ELF OpenGo is the first open-source Go
AI to convincingly demonstrate superhuman performance with a perfect (20:0) reco
rd against global top professionals. We apply ELF OpenGo to conduct extensive ab
lation studies, and to identify and analyze numerous interesting phenomena in bo
th the model training and in the gameplay inference procedures. Our code, models
, selfplay datasets, and auxiliary data are publicly available.
****
Random Matrix Improved Covariance Estimation for a Large Class of Metrics
Malik Tiomoko, Romain Couillet, Florent Bouchard, Guillaume Ginolhac
Relying on recent advances in statistical estimation of covariance distances bas
ed on random matrix theory, this article proposes an improved covariance and pre
cision matrix estimation for a wide family of metrics. The method is shown to la
rgely outperform the sample covariance matrix estimate and to compete with state
-of-the-art methods, while at the same time being computationally simpler and fa
ster. Applications to linear and quadratic discriminant analyses also show signi
ficant gains, therefore suggesting practical interest to statistical machine lea
rning.
****
Transfer of Samples in Policy Search via Multiple Importance Sampling
Andrea Tirinzoni, Mattia Salvini, Marcello Restelli
We consider the transfer of experience samples in reinforcement learning. Most o
f the previous works in this context focused on value-based settings, where tran
sferring instances conveniently reduces to the transfer of (s,a,s’,r) tuples. In
 this paper, we consider the more complex case of reusing samples in policy sear
ch methods, in which the agent is required to transfer entire trajectories betwe
en environments with different transition models. By leveraging ideas from multi
ple importance sampling, we propose robust gradient estimators that effectively
achieve this goal, along with several techniques to reduce their variance. In th
e case where the transition models are known, we theoretically establish the rob
ustness to the negative transfer for our estimators. In the case of unknown mode
ls, we propose a method to efficiently estimate them when the target task belong
s to a finite set of possible tasks and when it belongs to some reproducing kern
el Hilbert space. We provide empirical results to show the effectiveness of our
estimators.
****
Optimal Transport for structured data with application on graphs
Vayer Titouan, Nicolas Courty, Romain Tavenard, Chapel Laetitia, Rémi Flamary
This work considers the problem of computing distances between structured object
s such as undirected graphs, seen as probability distributions in a specific met
ric space. We consider a new transportation distance (i.e. that minimizes a tot
al cost of transporting probability masses) that unveils the geometric nature of
 the structured objects space. Unlike Wasserstein or Gromov-Wasserstein metrics
that focus solely and respectively on features (by considering a metric in the f
eature space) or structure (by seeing structure as a metric space), our new dist
ance exploits jointly both information, and is consequently called Fused Gromov-
Wasserstein (FGW). After discussing its properties and computational aspects, we
 show results on a graph classification task, where our method outperforms both
graph kernels and deep graph convolutional networks. Exploiting further on the m
etric properties of FGW, interesting geometric objects such as Fr{é}chet means o
r barycenters of graphs are illustrated and discussed in a clustering context.
****
Discovering Latent Covariance Structures for Multiple Time Series
Anh Tong, Jaesik Choi
Analyzing multivariate time series data is important to predict future events an
d changes of complex systems in finance, manufacturing, and administrative decis
ions. The expressiveness power of Gaussian Process (GP) regression methods has b
een significantly improved by compositional covariance structures. In this paper
, we present a new GP model which naturally handles multiple time series by plac

ing an Indian Buffet Process (IBP) prior on the presence of shared kernels. Our selective covariance structure decomposition allows exploiting shared parameters over a set of multiple, selected time series. We also investigate the well-definedness of the models when infinite latent components are introduced. We present a pragmatic search algorithm which explores a larger structure space efficiently. Experiments conducted on five real-world data sets demonstrate that our new model outperforms existing methods in term of structure discoveries and predictive performances.
****
Bayesian Generative Active Deep Learning
Toan Tran, Thanh-Toan Do, Ian Reid, Gustavo Carneiro
Deep learning models have demonstrated outstanding performance in several problems, but their training process tends to require immense amounts of computational and human resources for training and labeling, constraining the types of problems that can be tackled. Therefore, the design of effective training methods that require small labeled training sets is an important research direction that will allow a more effective use of resources. Among current approaches designed to address this issue, two are particularly interesting: data augmentation and active learning. Data augmentation achieves this goal by artificially generating new training points, while active learning relies on the selection of the “most informative” subset of unlabeled training samples to be labelled by an oracle. Although successful in practice, data augmentation can waste computational resources because it indiscriminately generates samples that are not guaranteed to be informative, and active learning selects a small subset of informative samples (from a large un-annotated set) that may be insufficient for the training process. In this paper, we propose a Bayesian generative active deep learning approach that combines active learning with data augmentation – we provide theoretical and empirical evidence (MNIST, CIFAR-$\{10,100\}$, and SVHN) that our approach has more efficient training and better classification results than data augmentation and active learning.
****
DeepNose: Using artificial neural networks to represent the space of odorants
Ngoc Tran, Daniel Kepple, Sergey Shuvaev, Alexei Koulakov
The olfactory system employs an ensemble of odorant receptors (ORs) to sense odorants and to derive olfactory percepts. We trained artificial neural networks to represent the chemical space of odorants and used this representation to predict human olfactory percepts. We hypothesized that ORs may be considered 3D convolutional filters that extract molecular features and, as such, can be trained using machine learning methods. First, we trained a convolutional autoencoder, called DeepNose, to deduce a low-dimensional representation of odorant molecules which were represented by their 3D spatial structure. Next, we tested the ability of DeepNose features in predicting physical properties and odorant percepts based on 3D molecular structure alone. We found that, despite the lack of human expertise, DeepNose features often outperformed molecular descriptors used in computational chemistry in predicting both physical properties and human perceptions. We propose that DeepNose network can extract de novo chemical features predictive of various bioactivities and can help understand the factors influencing the composition of ORs ensemble.
****
LR-GLM: High-Dimensional Bayesian Inference Using Low-Rank Data Approximations
Brian Trippe, Jonathan Huggins, Raj Agrawal, Tamara Broderick
Due to the ease of modern data collection, applied statisticians often have access to a large set of covariates that they wish to relate to some observed outcome. Generalized linear models (GLMs) offer a particularly interpretable framework for such an analysis. In these high-dimensional problems, the number of covariates is often large relative to the number of observations, so we face non-trivial inferential uncertainty; a Bayesian approach allows coherent quantification of this uncertainty. Unfortunately, existing methods for Bayesian inference in GLMs require running times roughly cubic in parameter dimension, and so are limited to settings with at most tens of thousand parameters. We propose to reduce time

and memory costs with a low-rank approximation of the data in an approach we call LR-GLM. When used with the Laplace approximation or Markov chain Monte Carlo, LR-GLM provides a full Bayesian posterior approximation and admits running times reduced by a full factor of the parameter dimension. We rigorously establish the quality of our approximation and show how the choice of rank allows a tunable computational–statistical trade-off. Experiments support our theory and demonstrate the efficacy of LR-GLM on real large-scale datasets.
****
Learning Hawkes Processes Under Synchronization Noise
William Trouleau, Jalal Etesami, Matthias Grossglauser, Negar Kiyavash, Patrick Thiran
Multivariate Hawkes processes (MHP) are widely used in a variety of fields to model the occurrence of discrete events. Prior work on learning MHPs has only focused on inference in the presence of perfect traces without noise. We address the problem of learning the causal structure of MHPs when observations are subject to an unknown delay. In particular, we introduce the so-called synchronization noise, where the stream of events generated by each dimension is subject to a random and unknown time shift. We characterize the robustness of the classic maximum likelihood estimator to synchronization noise, and we introduce a new approach for learning the causal structure in the presence of noise. Our experimental results show that our approach accurately recovers the causal structure of MHPs for a wide range of noise levels, and significantly outperforms classic estimation methods.
****
Homomorphic Sensing
Manolis Tsakiris, Liangzu Peng
A recent line of research termed "unlabeled sensing" and "shuffled linear regression" has been exploring under great generality the recovery of signals from subsampled and permuted measurements; a challenging problem in diverse fields of data science and machine learning. In this paper we introduce an abstraction of this problem which we call "homomorphic sensing". Given a linear subspace and a finite set of linear transformations we develop an algebraic theory which establishes conditions guaranteeing that points in the subspace are uniquely determined from their homomorphic image under some transformation in the set. As a special case, we recover known conditions for unlabeled sensing, as well as new results and extensions. On the algorithmic level we exhibit two dynamic programming based algorithms, which to the best of our knowledge are the first working solutions for the unlabeled sensing problem for small dimensions. One of them, additionally based on branch-and-bound, when applied to image registration under affine transformations, performs on par with or outperforms state-of-the-art methods on benchmark datasets.
****
Metropolis-Hastings Generative Adversarial Networks
Ryan Turner, Jane Hung, Eric Frank, Yunus Saatchi, Jason Yosinski
We introduce the Metropolis-Hastings generative adversarial network (MH-GAN), which combines aspects of Markov chain Monte Carlo and GANs. The MH-GAN draws samples from the distribution implicitly defined by a GAN’s discriminator-generator pair, as opposed to standard GANs which draw samples from the distribution defined only by the generator. It uses the discriminator from GAN training to build a wrapper around the generator for improved sampling. With a perfect discriminator, this wrapped generator samples from the true distribution on the data exactly even when the generator is imperfect. We demonstrate the benefits of the improved generator on multiple benchmark datasets, including CIFAR-10 and CelebA, using the DCGAN, WGAN, and progressive GAN.
****
Distributed, Egocentric Representations of Graphs for Detecting Critical Structures
Ruo-Chun Tzeng, Shan-Hung Wu
We study the problem of detecting critical structures using a graph embedding model. Existing graph embedding models lack the ability to precisely detect critic

al structures that are specific to a task at the global scale. In this paper, we
 propose a novel graph embedding model, called the Ego-CNNs, that employs the eg
o-convolutions convolutions at each layer and stacks up layers using an ego-cent
ric way to detects precise critical structures efficiently. An Ego-CNN can be jo
intly trained with a task model and help explain/discover knowledge for the task
. We conduct extensive experiments and the results show that Ego-CNNs (1) can le
ad to comparable task performance as the state-of-the-art graph embedding models
, (2) works nicely with CNN visualization techniques to illustrate the detected
structures, and (3) is efficient and can incorporate with scale-free priors, whi
ch commonly occurs in social network datasets, to further improve the training e
fficiency.
****
Sublinear Space Private Algorithms Under the Sliding Window Model
Jalaj Upadhyay
The Differential privacy overview of Apple states, “Apple retains the collected
data for a maximum of three months." Analysis of recent data is formalized by th
e sliding window model. This begs the question: what is the price of privacy in
the sliding window model? In this paper, we study heavy hitters in the sliding w
indow model with window size w. Previous works of Chan et al. (2012) estimates
 heavy hitters with an error of order θw for a constant $\theta >0$. In
this paper, we give an efficient differentially private algorithm to estimate he
avy hitters in the sliding window model with $\widetilde O(w^{3/4})$ additive er
ror and using $\widetilde O(\sqrt{w})$ space.
****
Fairness without Harm: Decoupled Classifiers with Preference Guarantees
Berk Ustun, Yang Liu, David Parkes
In domains such as medicine, it can be acceptable for machine learning models to
 include sensitive attributes such as gender and ethnicity. In this work, we arg
ue that when there is this kind of treatment disparity, then it should be in the
 best interest of each group. Drawing on ethical principles such as beneficence
("do the best") and non-maleficence ("do no harm"), we show how to use sensitive
 attributes to train decoupled classifiers that satisfy preference guarantees. T
hese guarantees ensure the majority of individuals in each group prefer their as
signed classifier to (i) a pooled model that ignores group membership (rationali
ty), and (ii) the model assigned to any other group (envy-freeness). We introduc
e a recursive procedure that adaptively selects group attributes for decoupling,
 and present formal conditions to ensure preference guarantees in terms of gener
alization error. We validate the effectiveness of the procedure on real-world da
tasets, showing that it improves accuracy without violating preference guarantee
s on test data.
****
Large-Scale Sparse Kernel Canonical Correlation Analysis
Viivi Uurtio, Sahely Bhadra, Juho Rousu
This paper presents gradKCCA, a large-scale sparse non-linear canonical correlat
ion method. Like Kernel Canonical Correlation Analysis (KCCA), our method finds
non-linear relations through kernel functions, but it does not rely on a kernel
matrix, a known bottleneck for scaling up kernel methods. gradKCCA corresponds t
o solving KCCA with the additional constraint that the canonical projection dire
ctions in the kernel-induced feature space have preimages in the original data s
pace. Firstly, this modification allows us to very efficiently maximize kernel c
anonical correlation through an alternating projected gradient algorithm working
 in the original data space. Secondly, we can control the sparsity of the projec
tion directions by constraining the ℓ_1 norm of the preimages of the projec
tion directions, facilitating the interpretation of the discovered patterns, whi
ch is not available through KCCA. Our empirical experiments demonstrate that gra
dKCCA outperforms state-of-the-art CCA methods in terms of speed and robustness
to noise both in simulated and real-world datasets.
****
Characterization of Convex Objective Functions and Optimal Expected Convergence
Rates for SGD

Marten Van Dijk, Lam Nguyen, Phuong Ha Nguyen, Dzung Phan
We study Stochastic Gradient Descent (SGD) with diminishing step sizes for conve x objective functions. We introduce a definitional framework and theory that def ines and characterizes a core property, called curvature, of convex objective fu nctions. In terms of curvature we can derive a new inequality that can be used t o compute an optimal sequence of diminishing step sizes by solving a differentia l equation. Our exact solutions confirm known results in literature and allows u s to fully characterize a new regularizer with its corresponding expected conver gence rates.
****
Composing Value Functions in Reinforcement Learning
Benjamin Van Niekerk, Steven James, Adam Earle, Benjamin Rosman
An important property for lifelong-learning agents is the ability to combine exi sting skills to solve new unseen tasks. In general, however, it is unclear how t o compose existing skills in a principled manner. Under the assumption of determ inistic dynamics, we prove that optimal value function composition can be achiev ed in entropy-regularised reinforcement learning (RL), and extend this result to the standard RL setting. Composition is demonstrated in a high-dimensional vide o game, where an agent with an existing library of skills is immediately able to solve new tasks without the need for further learning.
****
Model Comparison for Semantic Grouping
Francisco Vargas, Kamen Brestnichki, Nils Hammerla
We introduce a probabilistic framework for quantifying the semantic similarity b etween two groups of embeddings. We formulate the task of semantic similarity as a model comparison task in which we contrast a generative model which jointly m odels two sentences versus one that does not. We illustrate how this framework c an be used for the Semantic Textual Similarity tasks using clear assumptions abo ut how the embeddings of words are generated. We apply model comparison that uti lises information criteria to address some of the shortcomings of Bayesian model comparison, whilst still penalising model complexity. We achieve competitive re sults by applying the proposed framework with an appropriate choice of likelihoo d on the STS datasets.
****
Learning Dependency Structures for Weak Supervision Models
Paroma Varma, Frederic Sala, Ann He, Alexander Ratner, Christopher Re
Labeling training data is a key bottleneck in the modern machine learning pipeli ne. Recent weak supervision approaches combine labels from multiple noisy source s by estimating their accuracies without access to ground truth labels; however, estimating the dependencies among these sources is a critical challenge. We foc us on a robust PCA-based algorithm for learning these dependency structures, est ablish improved theoretical recovery rates, and outperform existing methods on v arious real-world tasks. Under certain conditions, we show that the amount of un labeled data needed can scale sublinearly or even logarithmically with the numbe r of sources m, improving over previous efforts that ignore the sparsity pattern in the dependency structure and scale linearly in m. We provide an information-theoretic lower bound on the minimum sample complexity of the weak supervision s etting. Our method outperforms weak supervision approaches that assume condition ally-independent sources by up to 4.64 F1 points and previous structure learning approaches by up to 4.41 F1 points on real-world relation extraction and image classification tasks.
****
Probabilistic Neural Symbolic Models for Interpretable Visual Question Answering
Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, Dev i Parikh
We propose a new class of probabilistic neural-symbolic models, that have symbol ic functional programs as a latent, stochastic variable. Instantiated in the con text of visual question answering, our probabilistic formulation offers two key conceptual advantages over prior neural-symbolic models for VQA. Firstly, the pr ograms generated by our model are more understandable while requiring less numbe

r of teaching examples. Secondly, we show that one can pose counterfactual scena
rios to the model, to probe its beliefs on the programs that could lead to a spe
cified answer given an image. Our results on the CLEVR and SHAPES datasets verif
y our hypotheses, showing that the model gets better program (and answer) predic
tion accuracy even in the low data regime, and allows one to probe the coherence
 and consistency of reasoning performed.
****
Manifold Mixup: Better Representations by Interpolating Hidden States
Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, Da
vid Lopez-Paz, Yoshua Bengio
Deep neural networks excel at learning the training data, but often provide inco
rrect and confident predictions when evaluated on slightly different test exampl
es. This includes distribution shifts, outliers, and adversarial examples. To ad
dress these issues, we propose \manifoldmixup{}, a simple regularizer that encou
rages neural networks to predict less confidently on interpolations of hidden re
presentations. \manifoldmixup{} leverages semantic interpolations as additional
training signal, obtaining neural networks with smoother decision boundaries at
multiple levels of representation. As a result, neural networks trained with \ma
nifoldmixup{} learn flatter class-representations, that is, with fewer direction
s of variance. We prove theory on why this flattening happens under ideal condit
ions, validate it empirically on practical situations, and connect it to the pre
vious works on information theory and generalization. In spite of incurring no s
ignificant computation and being implemented in a few lines of code, \manifoldmi
xup{} improves strong baselines in supervised learning, robustness to single-ste
p adversarial attacks, and test log-likelihood.
****
Maximum Likelihood Estimation for Learning Populations of Parameters
Ramya Korlakai Vinayak, Weihao Kong, Gregory Valiant, Sham Kakade
Consider a setting with N independent individuals, each with an unknown parame
ter, $p_i \in [0, 1]$ drawn from some unknown distribution $P^\star$. After obse
rving the outcomes of t independent Bernoulli trials, i.e., $X_i \sim \text{Bi
nomial}(t, p_i)$ per individual, our objective is to accurately estimate $P^\sta
r$ in the sparse regime, namely when $t \ll N$. This problem arises in numerous
domains, including the social sciences, psychology, health-care, and biology, wh
ere the size of the population under study is usually large yet the number of ob
servations per individual is often limited. Our main result shows that, in this
sparse regime where $t \ll N$, the maximum likelihood estimator (MLE) is both st
atistically minimax optimal and efficiently computable. Precisely, for sufficien
tly large N, the MLE achieves the information theoretic optimal error bound of
 $\mathcal{O}(\frac{1}{t})$ for $t < c\log{N}$, with regards to the earth mover’
s distance (between the estimated and true distributions). More generally, in an
 exponentially large interval of t beyond $c \log{N}$, the MLE achieves the mi
nimax error bound of $\mathcal{O}(\frac{1}{\sqrt{t\log N}})$. In contrast, regar
dless of how large N is, the naive "plug-in" estimator for this problem only a
chieves the sub-optimal error of $\Theta(\frac{1}{\sqrt{t}})$. Empirically, we a
lso demonstrate the MLE performs well on both synthetic as well as real datasets
.
****
Understanding Priors in Bayesian Neural Networks at the Unit Level
Mariia Vladimirova, Jakob Verbeek, Pablo Mesejo, Julyan Arbel
We investigate deep Bayesian neural networks with Gaussian priors on the weights
 and a class of ReLU-like nonlinearities. Bayesian neural networks with Gaussian
 priors are well known to induce an L2, “weight decay”, regularization. Our resu
lts indicate a more intricate regularization effect at the level of the unit act
ivations. Our main result establishes that the induced prior distribution on the
 units before and after activation becomes increasingly heavy-tailed with the de
pth of the layer. We show that first layer units are Gaussian, second layer unit
s are sub-exponential, and units in deeper layers are characterized by sub-Weibu
ll distributions. Our results provide new theoretical insight on deep Bayesian n
eural networks, which we corroborate with simulation experiments.

****
On the Design of Estimators for Bandit Off-Policy Evaluation
Nikos Vlassis, Aurelien Bibaut, Maria Dimakopoulou, Tony Jebara
Off-policy evaluation is the problem of estimating the value of a target policy using data collected under a different policy. Given a base estimator for bandit off-policy evaluation and a parametrized class of control variates, we address the problem of computing a control variate in that class that reduces the risk of the base estimator. We derive the population risk as a function of the class parameters and we establish conditions that guarantee risk improvement. We present our main results in the context of multi-armed bandits, and we propose a simple design for contextual bandits that gives rise to an estimator that is shown to perform well in multi-class cost-sensitive classification datasets.
****
Learning to select for a predefined ranking
Aleksei Ustimenko, Aleksandr Vorobev, Gleb Gusev, Pavel Serdyukov
In this paper, we formulate a novel problem of learning to select a set of items maximizing the quality of their ordered list, where the order is predefined by some explicit rule. Unlike the classic information retrieval problem, in our setting, the predefined order of items in the list may not correspond to their quality in general. For example, this is a dominant scenario in personalized news and social media feeds, where items are ordered by publication time in a user interface. We propose new theoretically grounded algorithms based on direct optimization of the resulting list quality. Our offline and online experiments with a large-scale product search engine demonstrate the overwhelming advantage of our methods over the baselines in terms of all key quality metrics.
****
On the Limitations of Representing Functions on Sets
Edward Wagstaff, Fabian Fuchs, Martin Engelcke, Ingmar Posner, Michael A. Osborne
Recent work on the representation of functions on sets has considered the use of summation in a latent space to enforce permutation invariance. In particular, it has been conjectured that the dimension of this latent space may remain fixed as the cardinality of the sets under consideration increases. However, we demonstrate that the analysis leading to this conjecture requires mappings which are highly discontinuous and argue that this is only of limited practical use. Motivated by this observation, we prove that an implementation of this model via continuous mappings (as provided by e.g. neural networks or Gaussian processes) actually imposes a constraint on the dimensionality of the latent space. Practical universal function representation for set inputs can only be achieved with a latent dimension at least the size of the maximum number of input elements.
****
Graph Convolutional Gaussian Processes
Ian Walker, Ben Glocker
We propose a novel Bayesian nonparametric method to learn translation-invariant relationships on non-Euclidean domains. The resulting graph convolutional Gaussian processes can be applied to problems in machine learning for which the input observations are functions with domains on general graphs. The structure of these models allows for high dimensional inputs while retaining expressibility, as is the case with convolutional neural networks. We present applications of graph convolutional Gaussian processes to images and triangular meshes, demonstrating their versatility and effectiveness, comparing favorably to existing methods, despite being relatively simple models.
****
Gaining Free or Low-Cost Interpretability with Interpretable Partial Substitute
Tong Wang
This work addresses the situation where a black-box model with good predictive performance is chosen over its interpretable competitors, and we show interpretability is still achievable in this case. Our solution is to find an interpretable substitute on a subset of data where the black-box model is overkill or nearly overkill while leaving the rest to the black-box. This transparency is obtained

at minimal cost or no cost of the predictive performance. Under this framework, we develop a Hybrid Rule Sets (HyRS) model that uses decision rules to capture the subspace of data where the rules are as accurate or almost as accurate as the black-box provided. To train a HyRS, we devise an efficient search algorithm that iteratively finds the optimal model and exploits theoretically grounded strategies to reduce computation. Our framework is agnostic to the black-box during training. Experiments on structured and text data show that HyRS obtains an effective trade-off between transparency and interpretability.
****
Convolutional Poisson Gamma Belief Network
Chaojie Wang, Bo Chen, Sucheng Xiao, Mingyuan Zhou
For text analysis, one often resorts to a lossy representation that either completely ignores word order or embeds each word as a low-dimensional dense feature vector. In this paper, we propose convolutional Poisson factor analysis (CPFA) that directly operates on a lossless representation that processes the words in each document as a sequence of high-dimensional one-hot vectors. To boost its performance, we further propose the convolutional Poisson gamma belief network (CPGBN) that couples CPFA with the gamma belief network via a novel probabilistic pooling layer. CPFA forms words into phrases and captures very specific phrase-level topics, and CPGBN further builds a hierarchy of increasingly more general phrase-level topics. For efficient inference, we develop both a Gibbs sampler and a Weibull distribution based convolutional variational auto-encoder. Experimental results demonstrate that CPGBN can extract high-quality text latent representations that capture the word order information, and hence can be leveraged as a building block to enrich a wide variety of existing latent variable models that ignore word order.
****
Differentially Private Empirical Risk Minimization with Non-convex Loss Functions
Di Wang, Changyou Chen, Jinhui Xu
We study the problem of Empirical Risk Minimization (ERM) with (smooth) non-convex loss functions under the differential-privacy (DP) model. Existing approaches for this problem mainly adopt gradient norms to measure the error, which in general cannot guarantee the quality of the solution. To address this issue, we first study the expected excess empirical (or population) risk, which was primarily used as the utility to measure the quality for convex loss functions. Specifically, we show that the excess empirical (or population) risk can be upper bounded by $\tilde{O}(\frac{d\log (1/\delta)}{\log n\epsilon^2})$ in the (ϵ, δ)-DP settings, where n is the data size and d is the dimensionality of the space. The $\frac{1}{\log n}$ term in the empirical risk bound can be further improved to $\frac{1}{n^{\Omega(1)}}$ (when d is a constant) by a highly non-trivial analysis on the time-average error. To obtain more efficient solutions, we also consider the connection between achieving differential privacy and finding approximate local minimum. Particularly, we show that when the size n is large enough, there are (ϵ, δ)-DP algorithms which can find an approximate local minimum of the empirical risk with high probability in both the constrained and non-constrained settings. These results indicate that one can escape saddle points privately.
****
Random Expert Distillation: Imitation Learning via Expert Policy Support Estimation
Ruohan Wang, Carlo Ciliberto, Pierluigi Vito Amadori, Yiannis Demiris
We consider the problem of imitation learning from a finite set of expert trajectories, without access to reinforcement signals. The classical approach of extracting the expert’s reward function via inverse reinforcement learning, followed by reinforcement learning is indirect and may be computationally expensive. Recent generative adversarial methods based on matching the policy distribution between the expert and the agent could be unstable during training. We propose a new framework for imitation learning by estimating the support of the expert policy to compute a fixed reward function, which allows us to re-frame imitation learn

ing within the standard reinforcement learning setting. We demonstrate the effic
acy of our reward function on both discrete and continuous domains, achieving co
mparable or better performance than the state of the art under different reinfor
cement learning algorithms.
****
SATNet: Bridging deep learning and logical reasoning using a differentiable sati
sfiability solver
Po-Wei Wang, Priya Donti, Bryan Wilder, Zico Kolter
Integrating logical reasoning within deep learning architectures has been a majo
r goal of modern AI systems. In this paper, we propose a new direction toward th
is goal by introducing a differentiable (smoothed) maximum satisfiability (MAXSA
T) solver that can be integrated into the loop of larger deep learning systems.
Our (approximate) solver is based upon a fast coordinate descent approach to sol
ving the semidefinite program (SDP) associated with the MAXSAT problem. We show
how to analytically differentiate through the solution to this SDP and efficient
ly solve the associated backward pass. We demonstrate that by integrating this s
olver into end-to-end learning systems, we can learn the logical structure of ch
allenging problems in a minimally supervised fashion. In particular, we show tha
t we can learn the parity function using single-bit supervision (a traditionally
 hard task for deep networks) and learn how to play 9x9 Sudoku solely from examp
les. We also solve a “visual Sudoku” problem that maps images of Sudoku puzzles
to their associated logical solutions by combining our MAXSAT solver with a trad
itional convolutional architecture. Our approach thus shows promise in integrati
ng logical structures within deep learning.
****
Improving Neural Language Modeling via Adversarial Training
Dilin Wang, Chengyue Gong, Qiang Liu
Recently, substantial progress has been made in language modeling by using deep
neural networks. However, in practice, large scale neural language models have b
een shown to be prone to overfitting. In this paper, we present a simple yet hig
hly effective adversarial training mechanism for regularizing neural language mo
dels. The idea is to introduce adversarial noise to the output embedding layer w
hile training the models. We show that the optimal adversarial noise yields a si
mple closed form solution, thus allowing us to develop a simple and time efficie
nt algorithm. Theoretically, we show that our adversarial mechanism effectively
encourages the diversity of the embedding vectors, helping to increase the robus
tness of models. Empirically, we show that our method improves on the single mod
el state-of-the-art results for language modeling on Penn Treebank (PTB) and Wik
itext-2, achieving test perplexity scores of 46.01 and 38.65, respectively. When
 applied to machine translation, our method improves over various transformer-ba
sed translation baselines in BLEU scores on the WMT14 English-German and IWSLT14
 German-English tasks.
****
EigenDamage: Structured Pruning in the Kronecker-Factored Eigenbasis
Chaoqi Wang, Roger Grosse, Sanja Fidler, Guodong Zhang
Reducing the test time resource requirements of a neural network while preservin
g test accuracy is crucial for running inference on resource-constrained devices
. To achieve this goal, we introduce a novel network reparameterization based on
 the Kronecker-factored eigenbasis (KFE), and then apply Hessian-based structure
d pruning methods in this basis. As opposed to existing Hessian-based pruning al
gorithms which do pruning in parameter coordinates, our method works in the KFE
where different weights are approximately independent, enabling accurate pruning
 and fast computation. We demonstrate empirically the effectiveness of the propo
sed method through extensive experiments. In particular, we highlight that the i
mprovements are especially significant for more challenging datasets and network
s. With negligible loss of accuracy, an iterative-pruning version gives a 10x re
duction in model size and a 8x reduction in FLOPs on wide ResNet32.
****
Nonlinear Stein Variational Gradient Descent for Learning Diversified Mixture Mo
dels

Dilin Wang, Qiang Liu
Diversification has been shown to be a powerful mechanism for learning robust mo
dels in non-convex settings. A notable example is learning mixture models, in wh
ich enforcing diversity between the different mixture components allows us to pr
event the model collapsing phenomenon and capture more patterns from the observe
d data. In this work, we present a variational approach for diversity-promoting
learning, which leverages the entropy functional as a natural mechanism for enfo
rcing diversity. We develop a simple and efficient functional gradient-based alg
orithm for optimizing the variational objective function, which provides a signi
ficant generalization of Stein variational gradient descent (SVGD). We test our
method on various challenging real world problems, including deep embedded clust
ering and deep anomaly detection. Empirical results show that our method provide
s an effective mechanism for diversity-promoting learning, achieving substantial
 improvement over existing methods.
****
On the Convergence and Robustness of Adversarial Training
Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, Quanquan Gu
Improving the robustness of deep neural networks (DNNs) to adversarial examples
is an important yet challenging problem for secure deep learning. Across existin
g defense techniques, adversarial training with Projected Gradient Decent (PGD)
is amongst the most effective. Adversarial training solves a min-max optimizatio
n problem, with the inner maximization generating adversarial examples by maximi
zing the classification loss, and the outer minimization finding model parameter
s by minimizing the loss on adversarial examples generated from the inner maximi
zation. A criterion that measures how well the inner maximization is solved is t
herefore crucial for adversarial training. In this paper, we propose such a crit
erion, namely First-Order Stationary Condition for constrained optimization (FOS
C), to quantitatively evaluate the convergence quality of adversarial examples f
ound in the inner maximization. With FOSC, we find that to ensure better robustn
ess, it is essential to use adversarial examples with better convergence quality
 at the later stages of training. Yet at the early stages, high convergence qual
ity adversarial examples are not necessary and may even lead to poor robustness.
 Based on these observations, we propose a dynamic training strategy to graduall
y increase the convergence quality of the generated adversarial examples, which
significantly improves the robustness of adversarial training. Our theoretical a
nd empirical results show the effectiveness of the proposed method.
****
State-Regularized Recurrent Neural Networks
Cheng Wang, Mathias Niepert
Recurrent neural networks are a widely used class of neural architectures with t
wo shortcomings. First, it is difficult to understand what exactly they learn. S
econd, they tend to work poorly on sequences requiring long-term memorization, d
espite having this capacity in principle. We aim to address both shortcomings wi
th a class of recurrent networks that use a stochastic state transition mechanis
m between cell applications. This mechanism, which we term state-regularization,
 makes RNNs transition between a finite set of learnable states. We evaluate sta
te-regularized RNNs on (1) regular languages for the purpose of automata extract
ion; (2) nonregular languages such as balanced parentheses, palindromes, and the
 copy task where external memory is required; and (3) real-word sequence learnin
g tasks for sentiment analysis, visual object recognition, and language modeling
. We show that state-regularization simplifies the extraction of finite state au
tomata from the RNN’s state transition dynamics; forces RNNs to operate more lik
e automata with external memory and less like finite state machines; and makes R
NNs more interpretable.
****
Deep Factors for Forecasting
Yuyang Wang, Alex Smola, Danielle Maddix, Jan Gasthaus, Dean Foster, Tim Janusch
owski
Producing probabilistic forecasts for large collections of similar and/or depend
ent time series is a practically highly relevant, yet challenging task. Classica

l time series models fail to capture complex patterns in the data and multivariate techniques struggle to scale to large problem sizes, but their reliance on strong structural assumptions makes them data-efficient and allows them to provide estimates of uncertainty. The converse is true for models based on deep neural networks, which can learn complex patterns and dependencies given enough data. In this paper, we propose a hybrid model that incorporates the benefits of both approaches. Our new method is data-driven and scalable via a latent, global, deep component. It also handles uncertainty through a local classical model. We provide both theoretical and empirical evidence for the soundness of our approach through a necessary and sufficient decomposition of exchangeable time series into a global and a local part and extensive experiments. Our experiments demonstrate the advantages of our model both in term of data efficiency and computational complexity.
****
Repairing without Retraining: Avoiding Disparate Impact with Counterfactual Distributions
Hao Wang, Berk Ustun, Flavio Calmon
When the performance of a machine learning model varies over groups defined by sensitive attributes (e.g., gender or ethnicity), the performance disparity can be expressed in terms of the probability distributions of the input and output variables over each group. In this paper, we exploit this fact to reduce the disparate impact of a fixed classification model over a population of interest. Given a black-box classifier, we aim to eliminate the performance gap by perturbing the distribution of input variables for the disadvantaged group. We refer to the perturbed distribution as a counterfactual distribution, and characterize its properties for common fairness criteria. We introduce a descent algorithm to learn a counterfactual distribution from data. We then discuss how the estimated distribution can be used to build a data preprocessor that can reduce disparate impact without training a new model. We validate our approach through experiments on real-world datasets, showing that it can repair different forms of disparity without a significant drop in accuracy.
****
On Sparse Linear Regression in the Local Differential Privacy Model
Di Wang, Jinhui Xu
In this paper, we study the sparse linear regression problem under the Local Differential Privacy (LDP) model. We first show that polynomial dependency on the dimensionality p of the space is unavoidable for the estimation error in both non-interactive and sequential interactive local models, if the privacy of the whole dataset needs to be preserved. Similar limitations also exist for other types of error measurements and in the relaxed local models. This indicates that differential privacy in high dimensional space is unlikely achievable for the problem. With the understanding of this limitation, we then present two algorithmic results. The first one is a sequential interactive LDP algorithm for the low dimensional sparse case, called Locally Differentially Private Iterative Hard Thresholding (LDP-IHT), which achieves a near optimal upper bound. This algorithm is actually rather general and can be used to solve quite a few other problems, such as (Local) DP-ERM with sparsity constraints and sparse regression with non-linear measurements. The second one is for the restricted (high dimensional) case where only the privacy of the responses (labels) needs to be preserved. For this case, we show that the optimal rate of the error estimation can be made logarithmically depending on p (i.e., $\log p$) in the local model, where an upper bound is obtained by a label-privacy version of LDP-IHT. Experiments on real world and synthetic datasets confirm our theoretical analysis.
****
Doubly Robust Joint Learning for Recommendation on Data Missing Not at Random
Xiaojie Wang, Rui Zhang, Yu Sun, Jianzhong Qi
In recommender systems, usually the ratings of a user to most items are missing and a critical problem is that the missing ratings are often missing not at random (MNAR) in reality. It is widely acknowledged that MNAR ratings make it difficult to accurately predict the ratings and unbiasedly estimate the performance of

rating prediction. Recent approaches use imputed errors to recover the predicti
on errors for missing ratings, or weight observed ratings with the propensities
of being observed. These approaches can still be severely biased in performance
estimation or suffer from the variance of the propensities. To overcome these li
mitations, we first propose an estimator that integrates the imputed errors and
propensities in a doubly robust way to obtain unbiased performance estimation an
d alleviate the effect of the propensity variance. To achieve good performance g
uarantees, based on this estimator, we propose joint learning of rating predicti
on and error imputation, which outperforms the state-of-the-art approaches on fo
ur real-world datasets.
****
On the Generalization Gap in Reparameterizable Reinforcement Learning
Huan Wang, Stephan Zheng, Caiming Xiong, Richard Socher
Understanding generalization in reinforcement learning (RL) is a significant cha
llenge, as many common assumptions of traditional supervised learning theory do
not apply. We focus on the special class of reparameterizable RL problems, where
the trajectory distribution can be decomposed using the reparametrization trick
. For this problem class, estimating the expected return is efficient and the tr
ajectory can be computed deterministically given peripheral random variables, wh
ich enables us to study reparametrizable RL using supervised learning and transf
er learning theory. Through these relationships, we derive guarantees on the gap
between the expected and empirical return for both intrinsic and external error
s, based on Rademacher complexity as well as the PAC-Bayes bound. Our bound sugg
ests the generalization capability of reparameterizable RL is related to multipl
e factors including “smoothness” of the environment transition, reward and agent
policy function class. We also empirically verify the relationship between the
generalization gap and these factors through simulations.
****
Bias Also Matters: Bias Attribution for Deep Neural Network Explanation
Shengjie Wang, Tianyi Zhou, Jeff Bilmes
The gradient of a deep neural network (DNN) w.r.t. the input provides informatio
n that can be used to explain the output prediction in terms of the input featur
es and has been widely studied to assist in interpreting DNNs. In a linear model
(i.e., g(x) = wx + b), the gradient corresponds to the weights w. Such a model
can reasonably locally-linearly approximate a smooth nonlinear DNN, and hence th
e weights of this local model are the gradient. The bias b, however, is usually
overlooked in attribution methods. In this paper, we observe that since the bias
in a DNN also has a non-negligible contribution to the correctness of predictio
ns, it can also play a significant role in understanding DNN behavior. We propos
e a backpropagation-type algorithm “bias back-propagation (BBp)” that starts at
the output layer and iteratively attributes the bias of each layer to its input
nodes as well as combining the resulting bias term of the previous layer. Togeth
er with the backpropagation of the gradient generating w, we can fully recover t
he locally linear model g(x) = wx + b. In experiments, we show that BBp can gene
rate complementary and highly interpretable explanations.
****
Jumpout : Improved Dropout for Deep Neural Networks with ReLUs
Shengjie Wang, Tianyi Zhou, Jeff Bilmes
We discuss three novel insights about dropout for DNNs with ReLUs: 1) dropout en
courages each local linear piece of a DNN to be trained on data points from near
by regions; 2) the same dropout rate results in different (effective) deactivati
on rates for layers with different portions of ReLU-deactivated neurons; and 3)
the rescaling factor of dropout causes a normalization inconsistency between tra
ining and test when used together with batch normalization. The above leads to t
hree simple but nontrivial modifications resulting in our method “jumpout.” Jump
out samples the dropout rate from a monotone decreasing distribution (e.g., the
right half of a Gaussian), so each local linear piece is trained, with high prob
ability, to work better for data points from nearby than more distant regions. J
umpout moreover adaptively normalizes the dropout rate at each layer and every t
raining batch, so the effective deactivation rate on the activated neurons is ke

pt the same. Furthermore, it rescales the outputs for a better trade-off that ke
eps both the variance and mean of neurons more consistent between training and t
est phases, thereby mitigating the incompatibility between dropout and batch nor
malization. Jumpout significantly improves the performance of different neural n
ets on CIFAR10, CIFAR100, Fashion-MNIST, STL10, SVHN, ImageNet-1k, etc., while i
ntroducing negligible additional memory and computation costs.
****
AdaGrad Stepsizes: Sharp Convergence Over Nonconvex Landscapes
Rachel Ward, Xiaoxia Wu, Leon Bottou
Adaptive gradient methods such as AdaGrad and its variants update the stepsize i
n stochastic gradient descent on the fly according to the gradients received alo
ng the way; such methods have gained widespread use in large-scale optimization
for their ability to converge robustly, without the need to fine-tune parameters
 such as the stepsize schedule. Yet, the theoretical guarantees to date for AdaG
rad are for online and convex optimization. We bridge this gap by providing stro
ng theoretical guarantees for the convergence of AdaGrad over smooth, nonconvex
landscapes. We show that the norm version of AdaGrad (AdaGrad-Norm) converges to
 a stationary point at the $\mathcal{O}(\log(N)/\sqrt{N})$ rate in the stochasti
c setting, and at the optimal $\mathcal{O}(1/N)$ rate in the batch (non-stochast
ic) setting – in this sense, our convergence guarantees are “sharp”. In particul
ar, both our theoretical results and extensive numerical experiments imply that
AdaGrad-Norm is robust to the unknown Lipschitz constant and level of stochastic
 noise on the gradient.
****
Generalized Linear Rule Models
Dennis Wei, Sanjeeb Dash, Tian Gao, Oktay Gunluk
This paper considers generalized linear models using rule-based features, also r
eferred to as rule ensembles, for regression and probabilistic classification. R
ules facilitate model interpretation while also capturing nonlinear dependences
and interactions. Our problem formulation accordingly trades off rule set comple
xity and prediction accuracy. Column generation is used to optimize over an expo
nentially large space of rules without pre-generating a large subset of candidat
es or greedily boosting rules one by one. The column generation subproblem is so
lved using either integer programming or a heuristic optimizing the same objecti
ve. In experiments involving logistic and linear regression, the proposed method
s obtain better accuracy-complexity trade-offs than existing rule ensemble algor
ithms. At one end of the trade-off, the methods are competitive with less interp
retable benchmark models.
****
On the statistical rate of nonlinear recovery in generative models with heavy-ta
iled data
Xiaohan Wei, Zhuoran Yang, Zhaoran Wang
We consider estimating a high-dimensional vector from non-linear measurements wh
ere the unknown vector is represented by a generative model $G:\mathbb{R}^k\righ
tarrow\mathbb{R}^d$ with $k\ll d$. Such a model poses structural priors on the u
nknown vector without having a dedicated basis, and in particular allows new and
 efficient approaches solving recovery problems with number of measurements far
less than the ambient dimension of the vector. While progresses have been made r
ecently regarding theoretical understandings on the linear Gaussian measurements
, much less is known when the model is possibly misspecified and the measurement
s are non-Gaussian. In this paper, we make a step towards such a direction by co
nsidering the scenario where the measurements are non-Gaussian, subject to possi
bly unknown nonlinear transformations and the responses are heavy-tailed. We the
n propose new estimators via score functions based on the first and second order
 Stein’s identity, and prove the sample size bound of $m=\mathcal{O}(k\varepsilo
n^{-2}\log(L/\varepsilon))$ achieving an ε error in the form of expo
nential concentration inequalities. Furthermore, for the special case of multi-l
ayer ReLU generative model, we improve the sample bound by a logarithm factor to
 $m=\mathcal{O}(k\varepsilon^{-2}\log(d))$, matching the state-of-art statistica
l rate in compressed sensing for estimating k-sparse vectors. On the technical

side, we develop new chaining methods bounding heavy-tailed processes, which could be of independent interest.
****
CapsAndRuns: An Improved Method for Approximately Optimal Algorithm Configuration
Gellert Weisz, Andras Gyorgy, Csaba Szepesvari
We consider the problem of configuring general-purpose solvers to run efficiently on problem instances drawn from an unknown distribution, a problem of major interest in solver autoconfiguration. Following previous work, we focus on designing algorithms that find a configuration with near-optimal expected capped runtime while doing the least amount of work, with the cap chosen in a configuration-specific way so that most instances are solved. In this paper we present a new algorithm, CapsAndRuns, which finds a near-optimal configuration while using time that scales (in a problem dependent way) with the optimal expected capped runtime, significantly strengthening previous results which could only guarantee a bound that scaled with the potentially much larger optimal expected uncapped runtime. The new algorithm is simpler and more intuitive than the previous methods: first it estimates the optimal runtime cap for each configuration, then it uses a Bernstein race to find a near optimal configuration given the caps. Experiments verify that our method can significantly outperform its competitors.
****
Non-Monotonic Sequential Text Generation
Sean Welleck, Kianté Brantley, Hal Daumé Iii, Kyunghyun Cho
Standard sequential generation methods assume a pre-specified generation order, such as text generation methods which generate words from left to right. In this work, we propose a framework for training models of text generation that operate in non-monotonic orders; the model directly learns good orders, without any additional annotation. Our framework operates by generating a word at an arbitrary position, and then recursively generating words to its left and then words to its right, yielding a binary tree. Learning is framed as imitation learning, including a coaching method which moves from imitating an oracle to reinforcing the policy’s own preferences. Experimental results demonstrate that using the proposed method, it is possible to learn policies which generate text without pre-specifying a generation order, while achieving competitive performance with conventional left-to-right generation.
****
PROVEN: Verifying Robustness of Neural Networks with a Probabilistic Approach
Lily Weng, Pin-Yu Chen, Lam Nguyen, Mark Squillante, Akhilan Boopathy, Ivan Oseledets, Luca Daniel
We propose a novel framework PROVEN to \textbf{PRO}babilistically \textbf{VE}rify \textbf{N}eural network’s robustness with statistical guarantees. PROVEN provides probability certificates of neural network robustness when the input perturbation follow distributional characterization. Notably, PROVEN is derived from current state-of-the-art worst-case neural network robustness verification frameworks, and therefore it can provide probability certificates with little computational overhead on top of existing methods such as Fast-Lin, CROWN and CNN-Cert. Experiments on small and large MNIST and CIFAR neural network models demonstrate our probabilistic approach can tighten up robustness certificate to around $1.8 \times$ and $3.5 \times$ with at least a $99.99%$ confidence compared with the worst-case robustness certificate by CROWN and CNN-Cert.
****
Learning deep kernels for exponential family densities
Li Wenliang, Danica J. Sutherland, Heiko Strathmann, Arthur Gretton
The kernel exponential family is a rich class of distributions, which can be fit efficiently and with statistical guarantees by score matching. Being required to choose a priori a simple kernel such as the Gaussian, however, limits its practical applicability. We provide a scheme for learning a kernel parameterized by a deep network, which can find complex location-dependent local features of the data geometry. This gives a very rich class of density models, capable of fitting complex structures on moderate-dimensional problems. Compared to deep density

models fit via maximum likelihood, our approach provides a complementary set of strengths and tradeoffs: in empirical studies, the former can yield higher likelihoods, whereas the latter gives better estimates of the gradient of the log density, the score, which describes the distribution’s shape.
****
Improving Model Selection by Employing the Test Data
Max Westphal, Werner Brannath
Model selection and evaluation are usually strictly separated by means of data splitting to enable an unbiased estimation and a simple statistical inference for the unknown generalization performance of the final prediction model. We investigate the properties of novel evaluation strategies, namely when the final model is selected based on empirical performances on the test data. To guard against selection induced overoptimism, we employ a parametric multiple test correction based on the approximate multivariate distribution of performance estimates. Our numerical experiments involve training common machine learning algorithms (EN, CART, SVM, XGB) on various artificial classification tasks. At its core, our proposed approach improves model selection in terms of the expected final model performance without introducing overoptimism. We furthermore observed a higher probability for a successful evaluation study, making it easier in practice to empirically demonstrate a sufficiently high predictive performance.
****
Automatic Classifiers as Scientific Instruments: One Step Further Away from Ground-Truth
Jacob Whitehill, Anand Ramakrishnan
Automatic machine learning-based detectors of various psychological and social phenomena (e.g., emotion, stress, engagement) have great potential to advance basic science. However, when a detector d is trained to approximate an existing measurement tool (e.g., a questionnaire, observation protocol), then care must be taken when interpreting measurements collected using d since they are one step further removed from the under- lying construct. We examine how the accuracy of d, as quantified by the correlation q of d’s out- puts with the ground-truth construct U, impacts the estimated correlation between U (e.g., stress) and some other phenomenon V (e.g., academic performance). In particular: (1) We show that if the true correlation between U and V is r, then the expected sample correlation, over all vectors T n whose correlation with U is q, is qr. (2) We derive a formula for the probability that the sample correlation (over n subjects) using d is positive given that the true correlation is negative (and vice-versa); this probability can be substantial (around 20 - 30%) for values of n and q that have been used in recent affective computing studies. (3) With the goal to reduce the variance of correlations estimated by an automatic detector, we show that training multiple neural networks d(1) , . . . , d(m) using different training architectures and hyperparameters for the same detection task provides only limited “coverage” of T^n.
****
Moment-Based Variational Inference for Markov Jump Processes
Christian Wildner, Heinz Koeppl
We propose moment-based variational inference as a flexible framework for approximate smoothing of latent Markov jump processes. The main ingredient of our approach is to partition the set of all transitions of the latent process into classes. This allows to express the Kullback-Leibler divergence from the approximate to the posterior process in terms of a set of moment functions that arise naturally from the chosen partition. To illustrate possible choices of the partition, we consider special classes of jump processes that frequently occur in applications. We then extend the results to latent parameter inference and demonstrate the method on several examples.
****
End-to-End Probabilistic Inference for Nonstationary Audio Analysis
William Wilkinson, Michael Andersen, Joshua D. Reiss, Dan Stowell, Arno Solin
A typical audio signal processing pipeline includes multiple disjoint analysis stages, including calculation of a time-frequency representation followed by spec

trogram-based feature analysis. We show how time-frequency analysis and nonnegat ive matrix factorisation can be jointly formulated as a spectral mixture Gaussia n process model with nonstationary priors over the amplitude variance parameters . Further, we formulate this nonlinear model's state space representation, makin g it amenable to infinite-horizon Gaussian process regression with approximate i nference via expectation propagation, which scales linearly in the number of tim e steps and quadratically in the state dimensionality. By doing so, we are able to process audio signals with hundreds of thousands of data points. We demonstra te, on various tasks with empirical data, how this inference scheme outperforms more standard techniques that rely on extended Kalman filtering.
****
Fairness risk measures
Robert Williamson, Aditya Menon
Ensuring that classifiers are non-discriminatory or fair with respect to a sensi tive feature (e.g., race or gender) is a topical problem. Progress in this task requires fixing a definition of fairness, and there have been several proposals in this regard over the past few years. Several of these, however, assume either binary sensitive features (thus precluding categorical or real-valued sensitive groups), or result in non-convex objectives (thus adversely affecting the optim isation landscape). In this paper, we propose a new definition of fairness that generalises some existing proposals, while allowing for generic sensitive featur es and resulting in a convex objective. The key idea is to enforce that the expe cted losses (or risks) across each subgroup induced by the sensitive feature are commensurate. We show how this relates to the rich literature on risk measures from mathematical finance. As a special case, this leads to a new convex fairnes s-aware objective based on minimising the conditional value at risk (CVaR).
****
Partially Exchangeable Networks and Architectures for Learning Summary Statistic s in Approximate Bayesian Computation
Samuel Wiqvist, Pierre-Alexandre Mattei, Umberto Picchini, Jes Frellsen
We present a novel family of deep neural architectures, named partially exchange able networks (PENs) that leverage probabilistic symmetries. By design, PENs are invariant to block-switch transformations, which characterize the partial excha ngeability properties of conditionally Markovian processes. Moreover, we show th at any block-switch invariant function has a PEN-like representation. The DeepSe ts architecture is a special case of PEN and we can therefore also target fully exchangeable data. We employ PENs to learn summary statistics in approximate Bay esian computation (ABC). When comparing PENs to previous deep learning methods f or learning summary statistics, our results are highly competitive, both conside ring time series and static models. Indeed, PENs provide more reliable posterior samples even when using less training data.
****
Wasserstein Adversarial Examples via Projected Sinkhorn Iterations
Eric Wong, Frank Schmidt, Zico Kolter
A rapidly growing area of work has studied the existence of adversarial examples , datapoints which have been perturbed to fool a classifier, but the vast majori ty of these works have focused primarily on threat models defined by ℓ_p no rm-bounded perturbations. In this paper, we propose a new threat model for adver sarial attacks based on the Wasserstein distance. In the image classification se tting, such distances measure the cost of moving pixel mass, which can naturally represent “standard” image manipulations such as scaling, rotation, translation , and distortion (and can potentially be applied to other settings as well). To generate Wasserstein adversarial examples, we develop a procedure for approximat e projection onto the Wasserstein ball, based upon a modified version of the Sin khorn iteration. The resulting algorithm can successfully attack image classific ation models, bringing traditional CIFAR10 models down to 3% accuracy within a W asserstein ball with radius 0.1 (i.e., moving 10% of the image mass 1 pixel), an d we demonstrate that PGD-based adversarial training can improve this adversaria l accuracy to 76%. In total, this work opens up a new direction of study in adve rsarial robustness, more formally considering convex metrics that accurately cap

ture the invariances that we typically believe should exist in classifiers, and code for all experiments in the paper is available at https://github.com/locusla b/projected_sinkhorn.
****
Imitation Learning from Imperfect Demonstration
Yueh-Hua Wu, Nontawat Charoenphakdee, Han Bao, Voot Tangkaratt, Masashi Sugiyama
Imitation learning (IL) aims to learn an optimal policy from demonstrations. How ever, such demonstrations are often imperfect since collecting optimal ones is c ostly. To effectively learn from imperfect demonstrations, we propose a novel ap proach that utilizes confidence scores, which describe the quality of demonstrat ions. More specifically, we propose two confidence-based IL methods, namely two-step importance weighting IL (2IWIL) and generative adversarial IL with imperfec t demonstration and confidence (IC-GAIL). We show that confidence scores given o nly to a small portion of sub-optimal demonstrations significantly improve the p erformance of IL both theoretically and empirically.
****
Learning a Compressed Sensing Measurement Matrix via Gradient Unrolling
Shanshan Wu, Alex Dimakis, Sujay Sanghavi, Felix Yu, Daniel Holtmann-Rice, Dmitr y Storcheus, Afshin Rostamizadeh, Sanjiv Kumar
Linear encoding of sparse vectors is widely popular, but is commonly data-indepe ndent – missing any possible extra (but a priori unknown) structure beyond spars ity. In this paper we present a new method to learn linear encoders that adapt t o data, while still performing well with the widely used ℓ_1 decoder. The c onvex ℓ_1 decoder prevents gradient propagation as needed in standard gradi ent-based training. Our method is based on the insight that unrolling the convex decoder into T projected subgradient steps can address this issue. Our method can be seen as a data-driven way to learn a compressed sensing measurement matr ix. We compare the empirical performance of 10 algorithms over 6 sparse datasets (3 synthetic and 3 real). Our experiments show that there is indeed additional structure beyond sparsity in the real datasets; our method is able to discover i t and exploit it to create excellent reconstructions with fewer measurements (by a factor of 1.1-3x) compared to the previous state-of-the-art methods. We illus trate an application of our method in learning label embeddings for extreme mult i-label classification, and empirically show that our method is able to match or outperform the precision scores of SLEEC, which is one of the state-of-the-art embedding-based approaches.
****
Heterogeneous Model Reuse via Optimizing Multiparty Multiclass Margin
Xi-Zhu Wu, Song Liu, Zhi-Hua Zhou
Nowadays, many problems require learning a model from data owned by different pa rticipants who are restricted to share their examples due to privacy concerns, w hich is referred to as multiparty learning in the literature. In conventional mu ltiparty learning, a global model is usually trained from scratch via a communic ation protocol, ignoring the fact that each party may already have a local model trained on her own dataset. In this paper, we define a multiparty multiclass ma rgin to measure the global behavior of a set of heterogeneous local models, and propose a general learning method called HMR (Heterogeneous Model Reuse) to opti mize the margin. Our method reuses local models to approximate a global model, e ven when data are non-i.i.d distributed among parties, by exchanging few example s under predefined budget. Experiments on synthetic and real-world data covering different multiparty scenarios show the effectiveness of our proposal.
****
Deep Compressed Sensing
Yan Wu, Mihaela Rosca, Timothy Lillicrap
Compressed sensing (CS) provides an elegant framework for recovering sparse sign als from compressed measurements. For example, CS can exploit the structure of n atural images and recover an image from only a few random measurements. Unlike p opular autoencoding models, reconstruction in CS is posed as an optimisation pro blem that is separate from sensing. CS is flexible and data efficient, but its a pplication has been restricted by the strong assumption of sparsity and costly r

econstruction process. A recent approach that combines CS with neural network generators has removed the constraint of sparsity, but reconstruction remains slow. Here we propose a novel framework that significantly improves both the performance and speed of signal recovery by jointly training a generator and the optimisation process for reconstruction via meta-learning. We explore training the measurements with different objectives, and derive a family of models based on minimising measurement errors. We show that Generative Adversarial Nets (GANs) can be viewed as a special case in this family of models. Borrowing insights from the CS perspective, we develop a novel way of improving GANs using gradient information from the discriminator.
****
Simplifying Graph Convolutional Networks
Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, Kilian Weinberger
Graph Convolutional Networks (GCNs) and their variants have experienced significant attention and have become the de facto methods for learning graph representations. GCNs derive inspiration primarily from recent deep learning approaches, and as a result, may inherit unnecessary complexity and redundant computation. In this paper, we reduce this excess complexity through successively removing nonlinearities and collapsing weight matrices between consecutive layers. We theoretically analyze the resulting linear model and show that it corresponds to a fixed low-pass filter followed by a linear classifier. Notably, our experimental evaluation demonstrates that these simplifications do not negatively impact accuracy in many downstream applications. Moreover, the resulting model scales to larger datasets, is naturally interpretable, and yields up to two orders of magnitude speedup over FastGCN.
****
Domain Adaptation with Asymmetrically-Relaxed Distribution Alignment
Yifan Wu, Ezra Winston, Divyansh Kaushik, Zachary Lipton
Domain adaptation addresses the common situation in which the target distribution generating our test data differs from the source distribution generating our training data. While absent assumptions, domain adaptation is impossible, strict conditions, e.g. covariate or label shift, enable principled algorithms. Recently-proposed domain-adversarial approaches consist of aligning source and target encodings, an approach often motivated as minimizing two (of three) terms in a theoretical bound on target error. Unfortunately, this minimization can cause arbitrary increases in the third term, a problem guaranteed to arise under shifting label distributions. We propose asymmetrically-relaxed distribution alignment, a new approach that overcomes some limitations of standard domain-adversarial algorithms. Moreover, we characterize precise assumptions under which our algorithm is theoretically principled and demonstrate empirical benefits on both synthetic and real datasets.
****
On Scalable and Efficient Computation of Large Scale Optimal Transport
Yujia Xie, Minshuo Chen, Haoming Jiang, Tuo Zhao, Hongyuan Zha
Optimal Transport (OT) naturally arises in many machine learning applications, yet the heavy computational burden limits its wide-spread uses. To address the scalability issue, we propose an implicit generative learning-based framework called SPOT (Scalable Push-forward of Optimal Transport). Specifically, we approximate the optimal transport plan by a pushforward of a reference distribution, and cast the optimal transport problem into a minimax problem. We then can solve OT problems efficiently using primal dual stochastic gradient-type algorithms. We also show that we can recover the density of the optimal transport plan using neural ordinary differential equations. Numerical experiments on both synthetic and real datasets illustrate that SPOT is robust and has favorable convergence behavior. SPOT also allows us to efficiently sample from the optimal transport plan, which benefits downstream applications such as domain adaptation.
****
Zeno: Distributed Stochastic Gradient Descent with Suspicion-based Fault-tolerance

Cong Xie, Sanmi Koyejo, Indranil Gupta
We present Zeno, a technique to make distributed machine learning, particularly
Stochastic Gradient Descent (SGD), tolerant to an arbitrary number of faulty wor
kers. Zeno generalizes previous results that assumed a majority of non-faulty no
des; we need assume only one non-faulty worker. Our key idea is to suspect worke
rs that are potentially defective. Since this is likely to lead to false positiv
es, we use a ranking-based preference mechanism. We prove the convergence of SGD
 for non-convex problems under these scenarios. Experimental results show that Z
eno outperforms existing approaches.
****
Differentiable Linearized ADMM
Xingyu Xie, Jianlong Wu, Guangcan Liu, Zhisheng Zhong, Zhouchen Lin
Recently, a number of learning-based optimization methods that combine data-driv
en architectures with the classical optimization algorithms have been proposed a
nd explored, showing superior empirical performance in solving various ill-posed
 inverse problems, but there is still a scarcity of rigorous analysis about the
convergence behaviors of learning-based optimization. In particular, most existi
ng analyses are specific to unconstrained problems but cannot apply to the more
general cases where some variables of interest are subject to certain constraint
s. In this paper, we propose Differentiable Linearized ADMM (D-LADMM) for solvin
g the problems with linear constraints. Specifically, D-LADMM is a K-layer LADMM
 inspired deep neural network, which is obtained by firstly introducing some lea
rnable weights in the classical Linearized ADMM algorithm and then generalizing
the proximal operator to some learnable activation function. Notably, we rigorou
sly prove that there exist a set of learnable parameters for D-LADMM to generate
 globally converged solutions, and we show that those desired parameters can be
attained by training D-LADMM in a proper way. To the best of our knowledge, we a
re the first to provide the convergence analysis for the learning-based optimiza
tion method on constrained problems.
****
Calibrated Approximate Bayesian Inference
Hanwen Xing, Geoff Nicholls, Jeong Lee
We give a general purpose computational framework for estimating the bias in cov
erage resulting from making approximations in Bayesian inference. Coverage is th
e probability credible sets cover true parameter values. We show how to estimate
 the actual coverage an approximation scheme achieves when the ideal observation
 model and the prior can be simulated, but have been replaced, in the Monte Carl
o, with approximations as they are intractable. Coverage estimation procedures g
iven in Lee et al. (2018) work well on simple problems, but are biased, and do n
ot scale well, as those authors note. For example, the methods of Lee et al. (20
18) fail for calibration of an approximate completely collapsed MCMC algorithm f
or partition structure in a Dirichlet process for clustering group labels in a h
ierarchical model. By exploiting the symmetry of the coverage error under permut
ation of low level group labels and smoothing with Bayesian Additive Regression
Trees, we are able to show that the original approximate inference had poor cove
rage and should not be trusted.
****
Power k-Means Clustering
Jason Xu, Kenneth Lange
Clustering is a fundamental task in unsupervised machine learning. Lloyd’s 1957
algorithm for k-means clustering remains one of the most widely used due to its
speed and simplicity, but the greedy approach is sensitive to initialization and
 often falls short at a poor solution. This paper explores an alternative to Llo
yd’s algorithm that retains its simplicity and mitigates its tendency to get tra
pped by local minima. Called power k-means, our method embeds the k-means proble
m in a continuous class of similar, better behaved problems with fewer local min
ima. Power k-means anneals its way toward the solution of ordinary k-means by wa
y of majorization-minimization (MM), sharing the appealing descent property and
low complexity of Lloyd’s algorithm. Further, our method complements widely used
 seeding strategies, reaping marked improvements when used together as demonstra

ted on a suite of simulated and real data examples.
****
Gromov-Wasserstein Learning for Graph Matching and Node Embedding
Hongteng Xu, Dixin Luo, Hongyuan Zha, Lawrence Carin Duke
A novel Gromov-Wasserstein learning framework is proposed to jointly match (alig
n) graphs and learn embedding vectors for the associated graph nodes. Using Grom
ov-Wasserstein discrepancy, we measure the dissimilarity between two graphs and
find their correspondence, according to the learned optimal transport. The node
embeddings associated with the two graphs are learned under the guidance of the
optimal transport, the distance of which not only reflects the topological struc
ture of each graph but also yields the correspondence across the graphs. These t
wo learning steps are mutually-beneficial, and are unified here by minimizing th
e Gromov-Wasserstein discrepancy with structural regularizers. This framework le
ads to an optimization problem that is solved by a proximal point method. We app
ly the proposed method to matching problems in real-world networks, and demonstr
ate its superior performance compared to alternative approaches.
****
Stochastic Optimization for DC Functions and Non-smooth Non-convex Regularizers
with Non-asymptotic Convergence
Yi Xu, Qi Qi, Qihang Lin, Rong Jin, Tianbao Yang
Difference of convex (DC) functions cover a broad family of non-convex and possi
bly non-smooth and non-differentiable functions, and have wide applications in m
achine learning and statistics. Although deterministic algorithms for DC functio
ns have been extensively studied, stochastic optimization that is more suitable
for learning with big data remains under-explored. In this paper, we propose new
 stochastic optimization algorithms and study their first-order convergence theo
ries for solving a broad family of DC functions. We improve the existing algorit
hms and theories of stochastic optimization for DC functions from both practical
 and theoretical perspectives. Moreover, we extend the proposed stochastic algor
ithms for DC functions to solve problems with a general non-convex non-different
iable regularizer, which does not necessarily have a DC decomposition but enjoys
 an efficient proximal mapping. To the best of our knowledge, this is the first
work that gives the first non-asymptotic convergence for solving non-convex opti
mization whose objective has a general non-convex non-differentiable regularizer
.
****
Learning a Prior over Intent via Meta-Inverse Reinforcement Learning
Kelvin Xu, Ellis Ratner, Anca Dragan, Sergey Levine, Chelsea Finn
A significant challenge for the practical application of reinforcement learning
to real world problems is the need to specify an oracle reward function that cor
rectly defines a task. Inverse reinforcement learning (IRL) seeks to avoid this
challenge by instead inferring a reward function from expert demonstrations. Whi
le appealing, it can be impractically expensive to collect datasets of demonstra
tions that cover the variation common in the real world (e.g. opening any type o
f door). Thus in practice, IRL must commonly be performed with only a limited se
t of demonstrations where it can be exceedingly difficult to unambiguously recov
er a reward function. In this work, we exploit the insight that demonstrations f
rom other tasks can be used to constrain the set of possible reward functions by
 learning a "prior" that is specifically optimized for the ability to infer expr
essive reward functions from limited numbers of demonstrations. We demonstrate t
hat our method can efficiently recover rewards from images for novel tasks and p
rovide intuition as to how our approach is analogous to learning a prior.
****
Variational Russian Roulette for Deep Bayesian Nonparametrics
Kai Xu, Akash Srivastava, Charles Sutton
Bayesian nonparametric models provide a principled way to automatically adapt th
e complexity of a model to the amount of the data available, but computation in
such models is difficult. Amortized variational approximations are appealing bec
ause of their computational efficiency, but current methods rely on a fixed fini
te truncation of the infinite model. This truncation level can be difficult to s

et, and also interacts poorly with amortized methods due to the over-pruning problem. Instead, we propose a new variational approximation, based on a method from statistical physics called Russian roulette sampling. This allows the variational distribution to adapt its complexity during inference, without relying on a fixed truncation level, and while still obtaining an unbiased estimate of the gradient of the original variational objective. We demonstrate this method on infinite sized variational auto-encoders using a Beta-Bernoulli (Indian buffet process) prior.
****

Supervised Hierarchical Clustering with Exponential Linkage
Nishant Yadav, Ari Kobren, Nicholas Monath, Andrew Mccallum
In supervised clustering, standard techniques for learning a pairwise dissimilarity function often suffer from a discrepancy between the training and clustering objectives, leading to poor cluster quality. Rectifying this discrepancy necessitates matching the procedure for training the dissimilarity function to the clustering algorithm. In this paper, we introduce a method for training the dissimilarity function in a way that is tightly coupled with hierarchical clustering, in particular single linkage. However, the appropriate clustering algorithm for a given dataset is often unknown. Thus we introduce an approach to supervised hierarchical clustering that smoothly interpolates between single, average, and complete linkage, and we give a training procedure that simultaneously learns a linkage function and a dissimilarity function. We accomplish this with a novel Exponential Linkage function that has a learnable parameter that controls the interpolation. In experiments on four datasets, our joint training procedure consistently matches or outperforms the next best training procedure/linkage function pair and gives up to 8 points improvement in dendrogram purity over discrepant pairs.
****

Learning to Prove Theorems via Interacting with Proof Assistants
Kaiyu Yang, Jia Deng
Humans prove theorems by relying on substantial high-level reasoning and problem-specific insights. Proof assistants offer a formalism that resembles human mathematical reasoning, representing theorems in higher-order logic and proofs as high-level tactics. However, human experts have to construct proofs manually by entering tactics into the proof assistant. In this paper, we study the problem of using machine learning to automate the interaction with proof assistants. We construct CoqGym, a large-scale dataset and learning environment containing 71K human-written proofs from 123 projects developed with the Coq proof assistant. We develop ASTactic, a deep learning-based model that generates tactics as programs in the form of abstract syntax trees (ASTs). Experiments show that ASTactic trained on CoqGym can generate effective tactics and can be used to prove new theorems not previously provable by automated methods. Code is available at https://github.com/princeton-vl/CoqGym.
****

Sample-Optimal Parametric Q-Learning Using Linearly Additive Features
Lin Yang, Mengdi Wang
Consider a Markov decision process (MDP) that admits a set of state-action features, which can linearly express the process’s probabilistic transition model. We propose a parametric Q-learning algorithm that finds an approximate-optimal policy using a sample size proportional to the feature dimension K and invariant with respect to the size of the state space. To further improve its sample efficiency, we exploit the monotonicity property and intrinsic noise structure of the Bellman operator, provided the existence of anchor state-actions that imply implicit non-negativity in the feature space. We augment the algorithm using techniques of variance reduction, monotonicity preservation, and confidence bounds. It is proved to find a policy which is ϵ-optimal from any initial state with high probability using $\widetilde{O}(K/\epsilon^2(1-\gamma)^3)$ sample transitions for arbitrarily large-scale MDP with a discount factor $\gamma\in(0,1)$. A matching information-theoretical lower bound is proved, confirming the sample optimality of the proposed method with respect to all parameters (up to polylog

factors).
****
LegoNet: Efficient Convolutional Neural Networks with Lego Filters
Zhaohui Yang, Yunhe Wang, Chuanjian Liu, Hanting Chen, Chunjing Xu, Boxin Shi, Chao Xu, Chang Xu
This paper aims to build efficient convolutional neural networks using a set of Lego filters. Many successful building blocks, e.g., inception and residual modules, have been designed to refresh state-of-the-art records of CNNs on visual recognition tasks. Beyond these high-level modules, we suggest that an ordinary filter in the neural network can be upgraded to a sophisticated module as well. Filter modules are established by assembling a shared set of Lego filters that are often of much lower dimensions. Weights in Lego filters and binary masks to stack Lego filters for these filter modules can be simultaneously optimized in an end-to-end manner as usual. Inspired by network engineering, we develop a split-transform-merge strategy for an efficient convolution by exploiting intermediate Lego feature maps. The compression and acceleration achieved by Lego Networks using the proposed Lego filters have been theoretically discussed. Experimental results on benchmark datasets and deep models demonstrate the advantages of the proposed Lego filters and their potential real-world applications on mobile devices.
****
SWALP : Stochastic Weight Averaging in Low Precision Training
Guandao Yang, Tianyi Zhang, Polina Kirichenko, Junwen Bai, Andrew Gordon Wilson, Chris De Sa
Low precision operations can provide scalability, memory savings, portability, and energy efficiency. This paper proposes SWALP, an approach to low precision training that averages low-precision SGD iterates with a modified learning rate schedule. SWALP is easy to implement and can match the performance of full-precision SGD even with all numbers quantized down to 8 bits, including the gradient accumulators. Additionally, we show that SWALP converges arbitrarily close to the optimal solution for quadratic objectives, and to a noise ball asymptotically smaller than low precision SGD in strongly convex settings.
****
ME-Net: Towards Effective Adversarial Robustness with Matrix Estimation
Yuzhe Yang, Guo Zhang, Dina Katabi, Zhi Xu
Deep neural networks are vulnerable to adversarial attacks. The literature is rich with algorithms that can easily craft successful adversarial examples. In contrast, the performance of defense techniques still lags behind. This paper proposes ME-Net, a defense method that leverages matrix estimation (ME). In ME-Net, images are preprocessed using two steps: first pixels are randomly dropped from the image; then, the image is reconstructed using ME. We show that this process destroys the adversarial structure of the noise, while re-enforcing the global structure in the original image. Since humans typically rely on such global structures in classifying images, the process makes the network mode compatible with human perception. We conduct comprehensive experiments on prevailing benchmarks such as MNIST, CIFAR-10, SVHN, and Tiny-ImageNet. Comparing ME-Net with state-of-the-art defense mechanisms shows that ME-Net consistently outperforms prior techniques, improving robustness against both black-box and white-box attacks.
****
Efficient Nonconvex Regularized Tensor Completion with Structure-aware Proximal Iterations
Quanming Yao, James Tin-Yau Kwok, Bo Han
Nonconvex regularizers have been successfully used in low-rank matrix learning. In this paper, we extend this to the more challenging problem of low-rank tensor completion. Based on the proximal average algorithm, we develop an efficient solver that avoids expensive tensor folding and unfolding. A special “sparse plus low-rank" structure, which is essential for fast computation of individual proximal steps, is maintained throughout the iterations. We also incorporate adaptive momentum to further speed up empirical convergence. Convergence results to critical points are provided under smoothness and Kurdyka-Lojasiewicz conditions. Ex

perimental results on a number of synthetic and real-world data sets show that t
he proposed algorithm is more efficient in both time and space, and is also more
 accurate than existing approaches.
****
Hierarchically Structured Meta-learning
Huaxiu Yao, Ying Wei, Junzhou Huang, Zhenhui Li
In order to learn quickly with few samples, meta-learning utilizes prior knowled
ge learned from previous tasks. However, a critical challenge in meta-learning i
s task uncertainty and heterogeneity, which can not be handled via globally shar
ing knowledge among tasks. In this paper, based on gradient-based meta-learning,
 we propose a hierarchically structured meta-learning (HSML) algorithm that expl
icitly tailors the transferable knowledge to different clusters of tasks. Inspir
ed by the way human beings organize knowledge, we resort to a hierarchical task
clustering structure to cluster tasks. As a result, the proposed approach not on
ly addresses the challenge via the knowledge customization to different clusters
 of tasks, but also preserves knowledge generalization among a cluster of simila
r tasks. To tackle the changing of task relationship, in addition, we extend the
 hierarchical structure to a continual learning environment. The experimental re
sults show that our approach can achieve state-of-the-art performance in both to
y-regression and few-shot image classification problems.
****
Tight Kernel Query Complexity of Kernel Ridge Regression and Kernel k-means Cl
ustering
Taisuke Yasuda, David Woodruff, Manuel Fernandez
Kernel methods generalize machine learning algorithms that only depend on the pa
irwise inner products of the dataset by replacing inner products with kernel eva
luations, a function that passes input points through a nonlinear feature map be
fore taking the inner product in a higher dimensional space. In this work, we pr
esent nearly tight lower bounds on the number of kernel evaluations required to
approximately solve kernel ridge regression (KRR) and kernel k-means clusterin
g (KKMC) on n input points. For KRR, our bound for relative error approximatio
n the argmin of the objective function is $\Omega(nd_{\mathrm{eff}}^\lambda/\var
epsilon)$ where d_{eff}^λ is the effective statistical dimensio
n, tight up to a $\log(d_{\mathrm{eff}}^\lambda/\varepsilon)$ factor. For KKMC,
our bound for finding a k-clustering achieving a relative error approximation
of the objective function is $\Omega(nk/\varepsilon)$, tight up to a $\log(k/\va
repsilon)$ factor. Our KRR result resolves a variant of an open question of El A
laoui and Mahoney, asking whether the effective statistical dimension is a lower
 bound on the sampling complexity or not. Furthermore, for the important input d
istribution case of mixtures of Gaussians, we provide algorithms that bypass the
 above lower bounds.
****
Understanding Geometry of Encoder-Decoder CNNs
Jong Chul Ye, Woon Kyoung Sung
Encoder-decoder networks using convolutional neural network (CNN) architecture h
ave been extensively used in deep learning literatures thanks to its excellent p
erformance for various inverse problems in computer vision, medical imaging, etc
. However, it is still difficult to obtain coherent geometric view why such an a
rchitecture gives the desired performance. Inspired by recent theoretical unders
tanding on generalizability, expressivity and optimization landscape of neural n
etworks, as well as the theory of convolutional framelets, here we provide a uni
fied theoretical framework that leads to a better understanding of geometry of e
ncoder-decoder CNNs. Our unified mathematical framework shows that encoder-decod
er CNN architecture is closely related to nonlinear basis representation using c
ombinatorial convolution frames, whose expressibility increases exponentially wi
th the network depth. We also demonstrate the importance of skipped connection i
n terms of expressibility, and optimization landscape.
****
Defending Against Saddle Point Attack in Byzantine-Robust Distributed Learning
Dong Yin, Yudong Chen, Ramchandran Kannan, Peter Bartlett

We study robust distributed learning that involves minimizing a non-convex loss function with saddle points. We consider the Byzantine setting where some worker machines have abnormal or even arbitrary and adversarial behavior, and in this setting, the Byzantine machines may create fake local minima near a saddle point that is far away from any true local minimum, even when robust gradient estimators are used. We develop ByzantinePGD, a robust first-order algorithm that can provably escape saddle points and fake local minima, and converge to an approximate true local minimizer with low iteration complexity. As a by-product, we give a simpler algorithm and analysis for escaping saddle points in the usual non-Byzantine setting. We further discuss three robust gradient estimators that can be used in ByzantinePGD, including median, trimmed mean, and iterative filtering. We characterize their performance in concrete statistical settings, and argue for their near-optimality in low and high dimensional regimes.
****
Rademacher Complexity for Adversarially Robust Generalization
Dong Yin, Ramchandran Kannan, Peter Bartlett
Many machine learning models are vulnerable to adversarial attacks; for example, adding adversarial perturbations that are imperceptible to humans can often make machine learning models produce wrong predictions with high confidence; moreover, although we may obtain robust models on the training dataset via adversarial training, in some problems the learned models cannot generalize well to the test data. In this paper, we focus on ℓ_∞ attacks, and study the adversarially robust generalization problem through the lens of Rademacher complexity. For binary linear classifiers, we prove tight bounds for the adversarial Rademacher complexity, and show that the adversarial Rademacher complexity is never smaller than its natural counterpart, and it has an unavoidable dimension dependence, unless the weight vector has bounded ℓ_1 norm, and our results also extend to multi-class linear classifiers; in addition, for (nonlinear) neural networks, we show that the dimension dependence in the adversarial Rademacher complexity also exists. We further consider a surrogate adversarial loss for one-hidden layer ReLU network and prove margin bounds for this setting. Our results indicate that having ℓ_1 norm constraints on the weight matrices might be a potential way to improve generalization in the adversarial setting. We demonstrate experimental results that validate our theoretical findings.
****
ARSM: Augment-REINFORCE-Swap-Merge Estimator for Gradient Backpropagation Through Categorical Variables
Mingzhang Yin, Yuguang Yue, Mingyuan Zhou
To address the challenge of backpropagating the gradient through categorical variables, we propose the augment-REINFORCE-swap-merge (ARSM) gradient estimator that is unbiased and has low variance. ARSM first uses variable augmentation, REINFORCE, and Rao-Blackwellization to re-express the gradient as an expectation under the Dirichlet distribution, then uses variable swapping to construct differently expressed but equivalent expectations, and finally shares common random numbers between these expectations to achieve significant variance reduction. Experimental results show ARSM closely resembles the performance of the true gradient for optimization in univariate settings; outperforms existing estimators by a large margin when applied to categorical variational auto-encoders; and provides a "try-and-see self-critic" variance reduction method for discrete-action policy gradient, which removes the need of estimating baselines by generating a random number of pseudo actions and estimating their action-value functions.
****
NAS-Bench-101: Towards Reproducible Neural Architecture Search
Chris Ying, Aaron Klein, Eric Christiansen, Esteban Real, Kevin Murphy, Frank Hutter
Recent advances in neural architecture search (NAS) demand tremendous computational resources, which makes it difficult to reproduce experiments and imposes a barrier-to-entry to researchers without access to large-scale computation. We aim to ameliorate these problems by introducing NAS-Bench-101, the first public architecture dataset for NAS research. To build NAS-Bench-101, we carefully constru

cted a compact, yet expressive, search space, exploiting graph isomorphisms to identify 423k unique convolutional architectures. We trained and evaluated all of these architectures multiple times on CIFAR-10 and compiled the results into a large dataset of over 5 million trained models. This allows researchers to evaluate the quality of a diverse range of models in milliseconds by querying the pre-computed dataset. We demonstrate its utility by analyzing the dataset as a whole and by benchmarking a range of architecture optimization algorithms.
****
TapNet: Neural Network Augmented with Task-Adaptive Projection for Few-Shot Learning
Sung Whan Yoon, Jun Seo, Jaekyun Moon
Handling previously unseen tasks after given only a few training examples continues to be a tough challenge in machine learning. We propose TapNets, neural networks augmented with task-adaptive projection for improved few-shot learning. Here, employing a meta-learning strategy with episode-based training, a network and a set of per-class reference vectors are learned across widely varying tasks. At the same time, for every episode, features in the embedding space are linearly projected into a new space as a form of quick task-specific conditioning. The training loss is obtained based on a distance metric between the query and the reference vectors in the projection space. Excellent generalization results in this way. When tested on the Omniglot, miniImageNet and tieredImageNet datasets, we obtain state of the art classification accuracies under various few-shot scenarios.
****
Towards Accurate Model Selection in Deep Unsupervised Domain Adaptation
Kaichao You, Ximei Wang, Mingsheng Long, Michael Jordan
Deep unsupervised domain adaptation (Deep UDA) methods successfully leverage rich labeled data in a source domain to boost the performance on related but unlabeled data in a target domain. However, algorithm comparison is cumbersome in Deep UDA due to the absence of accurate and standardized model selection method, posing an obstacle to further advances in the field. Existing model selection methods for Deep UDA are either highly biased, restricted, unstable, or even controversial (requiring labeled target data). To this end, we propose Deep Embedded Validation (DEV), which embeds adapted feature representation into the validation procedure to obtain unbiased estimation of the target risk with bounded variance. The variance is further reduced by the technique of control variate. The efficacy of the method has been justified both theoretically and empirically.
****
Position-aware Graph Neural Networks
Jiaxuan You, Rex Ying, Jure Leskovec
Learning node embeddings that capture a node’s position within the broader graph structure is crucial for many prediction tasks on graphs. However, existing Graph Neural Network (GNN) architectures have limited power in capturing the position/location of a given node with respect to all other nodes of the graph. Here we propose Position-aware Graph Neural Networks (P-GNNs), a new class of GNNs for computing position-aware node embeddings. P-GNN first samples sets of anchor nodes, computes the distance of a given target node to each anchor-set, and then learns a non-linear distance-weighted aggregation scheme over the anchor-sets. This way P-GNNs can capture positions/locations of nodes with respect to the anchor nodes. P-GNNs have several advantages: they are inductive, scalable, and can incorporate node feature information. We apply P-GNNs to multiple prediction tasks including link prediction and community detection. We show that P-GNNs consistently outperform state of the art GNNs, with up to 66% improvement in terms of the ROC AUC score.
****
Learning Neurosymbolic Generative Models via Program Synthesis
Halley Young, Osbert Bastani, Mayur Naik
Generative models have become significantly more powerful in recent years. However, these models continue to have difficulty capturing global structure in data. For example, images of buildings typically contain spatial patterns such as win

dows repeating at regular intervals, but state-of-the-art models have difficulty generating these patterns. We propose to address this problem by incorporating programs representing global structure into generative models{—}e.g., a 2D for-loop may represent a repeating pattern of windows{—}along with a framework for learning these models by leveraging program synthesis to obtain training data. On both synthetic and real-world data, we demonstrate that our approach substantially outperforms state-of-the-art at both generating and completing images with global structure.
****
DAG-GNN: DAG Structure Learning with Graph Neural Networks
Yue Yu, Jie Chen, Tian Gao, Mo Yu
Learning a faithful directed acyclic graph (DAG) from samples of a joint distribution is a challenging combinatorial problem, owing to the intractable search space superexponential in the number of graph nodes. A recent breakthrough formulates the problem as a continuous optimization with a structural constraint that ensures acyclicity (Zheng et al., 2018). The authors apply the approach to the linear structural equation model (SEM) and the least-squares loss function that are statistically well justified but nevertheless limited. Motivated by the widespread success of deep learning that is capable of capturing complex nonlinear mappings, in this work we propose a deep generative model and apply a variant of the structural constraint to learn the DAG. At the heart of the generative model is a variational autoencoder parameterized by a novel graph neural network architecture, which we coin DAG-GNN. In addition to the richer capacity, an advantage of the proposed model is that it naturally handles discrete variables as well as vector-valued ones. We demonstrate that on synthetic data sets, the proposed method learns more accurate graphs for nonlinearly generated samples; and on benchmark data sets with discrete variables, the learned graphs are reasonably close to the global optima. The code is available at \url{https://github.com/fishmoon1234/DAG-GNN}.
****
How does Disagreement Help Generalization against Label Corruption?
Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, Masashi Sugiyama
Learning with noisy labels is one of the hottest problems in weakly-supervised learning. Based on memorization effects of deep neural networks, training on small-loss instances becomes very promising for handling noisy labels. This fosters the state-of-the-art approach "Co-teaching" that cross-trains two deep neural networks using the small-loss trick. However, with the increase of epochs, two networks converge to a consensus and Co-teaching reduces to the self-training MentorNet. To tackle this issue, we propose a robust learning paradigm called Co-teaching+, which bridges the "Update by Disagreement” strategy with the original Co-teaching. First, two networks feed forward and predict all data, but keep prediction disagreement data only. Then, among such disagreement data, each network selects its small-loss data, but back propagates the small-loss data from its peer network and updates its own parameters. Empirical results on benchmark datasets demonstrate that Co-teaching+ is much superior to many state-of-the-art methods in the robustness of trained models.
****
On the Computation and Communication Complexity of Parallel SGD with Dynamic Batch Sizes for Stochastic Non-Convex Optimization
Hao Yu, Rong Jin
For SGD based distributed stochastic optimization, computation complexity, measured by the convergence rate in terms of the number of stochastic gradient calls, and communication complexity, measured by the number of inter-node communication rounds, are two most important performance metrics. The classical data-parallel implementation of SGD over N workers can achieve linear speedup of its convergence rate but incurs an inter-node communication round at each batch. We study the benefit of using dynamically increasing batch sizes in parallel SGD for stochastic non-convex optimization by charactering the attained convergence rate and the required number of communication rounds. We show that for stochastic non-convex optimization under the P-L condition, the classical data-parallel SGD with e

xponentially increasing batch sizes can achieve the fastest known $O(1/(NT))$ convergence with linear speedup using only $\log(T)$ communication rounds. For general stochastic non-convex optimization, we propose a Catalyst-like algorithm to achieve the fastest known $O(1/\sqrt{NT})$ convergence with only $O(\sqrt{NT}\log(\frac{T}{N}))$ communication rounds.
****
On the Linear Speedup Analysis of Communication Efficient Momentum SGD for Distributed Non-Convex Optimization
Hao Yu, Rong Jin, Sen Yang
Recent developments on large-scale distributed machine learning applications, e.g., deep neural networks, benefit enormously from the advances in distributed non-convex optimization techniques, e.g., distributed Stochastic Gradient Descent (SGD). A series of recent works study the linear speedup property of distributed SGD variants with reduced communication. The linear speedup property enables us to scale out the computing capability by adding more computing nodes into our system. The reduced communication complexity is desirable since communication overhead is often the performance bottleneck in distributed systems. Recently, momentum methods are more and more widely adopted by practitioners to train machine learning models since they can often converge faster and generalize better. However, it remains unclear whether any distributed momentum SGD possesses the same linear speedup property as distributed SGD and has reduced communication complexity. This paper fills the gap by considering a distributed communication efficient momentum SGD method and proving its linear speedup property.
****
Multi-Agent Adversarial Inverse Reinforcement Learning
Lantao Yu, Jiaming Song, Stefano Ermon
Reinforcement learning agents are prone to undesired behaviors due to reward mis-specification. Finding a set of reward functions to properly guide agent behaviors is particularly challenging in multi-agent scenarios. Inverse reinforcement learning provides a framework to automatically acquire suitable reward functions from expert demonstrations. Its extension to multi-agent settings, however, is difficult due to the more complex notions of rational behaviors. In this paper, we propose MA-AIRL, a new framework for multi-agent inverse reinforcement learning, which is effective and scalable for Markov games with high-dimensional state-action space and unknown dynamics. We derive our algorithm based on a new solution concept and maximum pseudolikelihood estimation within an adversarial reward learning framework. In the experiments, we demonstrate that MA-AIRL can recover reward functions that are highly correlated with the ground truth rewards, while significantly outperforms prior methods in terms of policy imitation.
****
Distributed Learning over Unreliable Networks
Chen Yu, Hanlin Tang, Cedric Renggli, Simon Kassing, Ankit Singla, Dan Alistarh, Ce Zhang, Ji Liu
Most of today’s distributed machine learning systems assume reliable networks: whenever two machines exchange information (e.g., gradients or models), the network should guarantee the delivery of the message. At the same time, recent work exhibits the impressive tolerance of machine learning algorithms to errors or noise arising from relaxed communication or synchronization. In this paper, we connect these two trends, and consider the following question: Can we design machine learning systems that are tolerant to network unreliability during training? With this motivation, we focus on a theoretical problem of independent interest—given a standard distributed parameter server architecture, if every communication between the worker and the server has a non-zero probability p of being dropped, does there exist an algorithm that still converges, and at what speed? In the context of prior art, this problem can be phrased as distributed learning over random topologies. The technical contribution of this paper is a novel theoretical analysis proving that distributed learning over random topologies can achieve comparable convergence rate to centralized or distributed learning over reliable networks. Further, we prove that the influence of the packet drop rate diminishes with the growth of the number of parameter servers. We map this theoretical

result onto a real-world scenario, training deep neural networks over an unreliable network layer, and conduct network simulation to validate the system improvement by allowing the networks to be unreliable.
****
Online Adaptive Principal Component Analysis and Its extensions
Jianjun Yuan, Andrew Lamperski
We propose algorithms for online principal component analysis (PCA) and variance minimization for adaptive settings. Previous literature has focused on upper bounding the static adversarial regret, whose comparator is the optimal fixed action in hindsight. However, static regret is not an appropriate metric when the underlying environment is changing. Instead, we adopt the adaptive regret metric from the previous literature and propose online adaptive algorithms for PCA and variance minimization, that have sub-linear adaptive regret guarantees. We demonstrate both theoretically and experimentally that the proposed algorithms can adapt to the changing environments.
****
Generative Modeling of Infinite Occluded Objects for Compositional Scene Representation
Jinyang Yuan, Bin Li, Xiangyang Xue
We present a deep generative model which explicitly models object occlusions for compositional scene representation. Latent representations of objects are disentangled into location, size, shape, and appearance, and the visual scene can be generated compositionally by integrating these representations and an infinite-dimensional binary vector indicating presences of objects in the scene. By training the model to learn spatial dependences of pixels in the unsupervised setting, the number of objects, pixel-level segregation of objects, and presences of objects in overlapping regions can be estimated through inference of latent variables. Extensive experiments conducted on a series of specially designed datasets demonstrate that the proposed method outperforms two state-of-the-art methods when object occlusions exist.
****
Differential Inclusions for Modeling Nonsmooth ADMM Variants: A Continuous Limit Theory
Huizhuo Yuan, Yuren Zhou, Chris Junchi Li, Qingyun Sun
Recently, there has been a great deal of research attention on understanding the convergence behavior of first-order methods. One line of this research focuses on analyzing the convergence behavior of first-order methods using tools from continuous dynamical systems such as ordinary differential equations and differential inclusions. These research results shed lights on better understanding first-order methods from a non-optimization point of view. The alternating direction method of multipliers (ADMM) is a widely used first-order method for solving optimization problems arising from machine learning and statistics, and it is important to investigate its behavior using these new techniques from dynamical systems. Existing works along this line have been mainly focusing on problems with smooth objective functions, which exclude many important applications that are traditionally solved by ADMM variants. In this paper, we analyze some well-known and widely used ADMM variants for nonsmooth optimization problems using tools of differential inclusions. In particular, we analyze the convergence behavior of linearized ADMM, gradient-based ADMM, generalized ADMM and accelerated generalized ADMM for nonsmooth problems and show their connections with dynamical systems. We anticipate that these results will provide new insights on understanding ADMM for solving nonsmooth problems.
****
Trimming the ℓ_1 Regularizer: Statistical Analysis, Optimization, and Applications to Deep Learning
Jihun Yun, Peng Zheng, Eunho Yang, Aurelie Lozano, Aleksandr Aravkin
We study high-dimensional estimators with the trimmed ℓ_1 penalty, which leaves the h largest parameter entries penalty-free. While optimization techniques for this nonconvex penalty have been studied, the statistical properties have not yet been analyzed. We present the first statistical analyses for M-estimation

, and characterize support recovery, ℓ_∞ and ℓ_2 error of the trimmed ℓ_1 estimates as a function of the trimming parameter h. Our results show different regimes based on how h compares to the true support size. Our second contribution is a new algorithm for the trimmed regularization problem, which has the same theoretical convergence rate as difference of convex (DC) algorithms, but in practice is faster and finds lower objective values. Empirical evaluation of ℓ_1 trimming for sparse linear regression and graphical model estimation indicate that trimmed ℓ_1 can outperform vanilla ℓ_1 and non-convex alternatives. Our last contribution is to show that the trimmed penalty is beneficial beyond M-estimation, and yields promising results for two deep learning tasks: input structures recovery and network sparsification.
****
Bayesian Nonparametric Federated Learning of Neural Networks
Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, Yasaman Khazaeni
In federated learning problems, data is scattered across different servers and exchanging or pooling it is often impractical or prohibited. We develop a Bayesian nonparametric framework for federated learning with neural networks. Each data server is assumed to provide local neural network weights, which are modeled through our framework. We then develop an inference approach that allows us to synthesize a more expressive global network without additional supervision, data pooling and with as few as a single communication round. We then demonstrate the efficacy of our approach on federated learning problems simulated from two popular image classification datasets.
****
Dirichlet Simplex Nest and Geometric Inference
Mikhail Yurochkin, Aritra Guha, Yuekai Sun, Xuanlong Nguyen
We propose Dirichlet Simplex Nest, a class of probabilistic models suitable for a variety of data types, and develop fast and provably accurate inference algorithms by accounting for the model’s convex geometry and low dimensional simplicial structure. By exploiting the connection to Voronoi tessellation and properties of Dirichlet distribution, the proposed inference algorithm is shown to achieve consistency and strong error bound guarantees on a range of model settings and data distributions. The effectiveness of our model and the learning algorithm is demonstrated by simulations and by analyses of text and financial data.
****
A Conditional-Gradient-Based Augmented Lagrangian Framework
Alp Yurtsever, Olivier Fercoq, Volkan Cevher
This paper considers a generic convex minimization template with affine constraints over a compact domain, which covers key semidefinite programming applications. The existing conditional gradient methods either do not apply to our template or are too slow in practice. To this end, we propose a new conditional gradient method, based on a unified treatment of smoothing and augmented Lagrangian frameworks. The proposed method maintains favorable properties of the classical conditional gradient method, such as cheap linear minimization oracle calls and sparse representation of the decision variable. We prove $O(1/\sqrt{k})$ convergence rate for our method in the objective residual and the feasibility gap. This rate is essentially the same as the state of the art CG-type methods for our problem template, but the proposed method is arguably superior in practice compared to existing methods in various applications.
****
Conditional Gradient Methods via Stochastic Path-Integrated Differential Estimator
Alp Yurtsever, Suvrit Sra, Volkan Cevher
We propose a class of variance-reduced stochastic conditional gradient methods. By adopting the recent stochastic path-integrated differential estimator technique (SPIDER) of Fang et. al. (2018) for the classical Frank-Wolfe (FW) method, we introduce SPIDER-FW for finite-sum minimization as well as the more general expectation minimization problems. SPIDER-FW enjoys superior complexity guarantees in the non-convex setting, while matching the best known FW variants in the conv

ex case. We also extend our framework a la conditional gradient sliding (CGS) of Lan & Zhou. (2016), and propose SPIDER-CGS.
****
Context-Aware Zero-Shot Learning for Object Recognition
Eloi Zablocki, Patrick Bordes, Laure Soulier, Benjamin Piwowarski, Patrick Gallinari
Zero-Shot Learning (ZSL) aims at classifying unlabeled objects by leveraging auxiliary knowledge, such as semantic representations. A limitation of previous approaches is that only intrinsic properties of objects, e.g. their visual appearance, are taken into account while their context, e.g. the surrounding objects in the image, is ignored. Following the intuitive principle that objects tend to be found in certain contexts but not others, we propose a new and challenging approach, context-aware ZSL, that leverages semantic representations in a new way to model the conditional likelihood of an object to appear in a given context. Finally, through extensive experiments conducted on Visual Genome, we show that contextual information can substantially improve the standard ZSL approach and is robust to unbalanced classes.
****
Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds
Andrea Zanette, Emma Brunskill
Strong worst-case performance bounds for episodic reinforcement learning exist but fortunately in practice RL algorithms perform much better than such bounds would predict. Algorithms and theory that provide strong problem-dependent bounds could help illuminate the key features of what makes a RL problem hard and reduce the barrier to using RL algorithms in practice. As a step towards this we derive an algorithm and analysis for finite horizon discrete MDPs with state-of-the-art worst-case regret bounds and substantially tighter bounds if the RL environment has special features but without apriori knowledge of the environment from the algorithm. As a result of our analysis, we also help address an open learning theory question \cite{jiang2018open} about episodic MDPs with a constant upper-bound on the sum of rewards, providing a regret bound function of the number of episodes with no dependence on the horizon.
****
Global Convergence of Block Coordinate Descent in Deep Learning
Jinshan Zeng, Tim Tsz-Kit Lau, Shaobo Lin, Yuan Yao
Deep learning has aroused extensive attention due to its great empirical success. The efficiency of the block coordinate descent (BCD) methods has been recently demonstrated in deep neural network (DNN) training. However, theoretical studies on their convergence properties are limited due to the highly nonconvex nature of DNN training. In this paper, we aim at providing a general methodology for provable convergence guarantees for this type of methods. In particular, for most of the commonly used DNN training models involving both two- and three-splitting schemes, we establish the global convergence to a critical point at a rate of ${\cal O}(1/k)$, where k is the number of iterations. The results extend to general loss functions which have Lipschitz continuous gradients and deep residual networks (ResNets). Our key development adds several new elements to the Kurdyka-Lojasiewicz inequality framework that enables us to carry out the global convergence analysis of BCD in the general scenario of deep learning.
****
Making Convolutional Networks Shift-Invariant Again
Richard Zhang
Modern convolutional networks are not shift-invariant, as small input shifts or translations can cause drastic changes in the output. Commonly used downsampling methods, such as max-pooling, strided-convolution, and average-pooling, ignore the sampling theorem. The well-known signal processing fix is anti-aliasing by low-pass filtering before downsampling. However, simply inserting this module into deep networks leads to performance degradation; as a result, it is seldomly used today. We show that when integrated correctly, it is compatible with existing architectural components, such as max-pooling. The technique is general and can

be incorporated across layer types and applications, such as image classification and conditional image generation. In addition to increased shift-invariance, we also observe, surprisingly, that anti-aliasing boosts accuracy in ImageNet classification, across several commonly-used architectures. This indicates that anti-aliasing serves as effective regularization. Our results demonstrate that this classical signal processing technique has been undeservingly overlooked in modern deep networks.
****
Warm-starting Contextual Bandits: Robustly Combining Supervised and Bandit Feedback
Chicheng Zhang, Alekh Agarwal, Hal Daumé Iii, John Langford, Sahand Negahban
We investigate the feasibility of learning from both fully-labeled supervised data and contextual bandit data. We specifically consider settings in which the underlying learning signal may be different between these two data sources. Theoretically, we state and prove no-regret algorithms for learning that is robust to divergences between the two sources. Empirically, we evaluate some of these algorithms on a large selection of datasets, showing that our approaches are feasible, and helpful in practice.
****
When Samples Are Strategically Selected
Hanrui Zhang, Yu Cheng, Vincent Conitzer
In standard classification problems, the assumption is that the entity making the decision (the principal) has access to all the samples. However, in many contexts, she either does not have direct access to the samples, or can inspect only a limited set of samples and does not know which are the most relevant ones. In such cases, she must rely on another party (the agent) to either provide the samples or point out the most relevant ones. If the agent has a different objective, then the principal cannot trust the submitted samples to be representative. She must set a policy for how she makes decisions, keeping in mind the agent's incentives. In this paper, we introduce a theoretical framework for this problem and provide key structural and computational results.
****
Self-Attention Generative Adversarial Networks
Han Zhang, Ian Goodfellow, Dimitris Metaxas, Augustus Odena
In this paper, we propose the Self-Attention Generative Adversarial Network (SAGAN) which allows attention-driven, long-range dependency modeling for image generation tasks. Traditional convolutional GANs generate high-resolution details as a function of only spatially local points in lower-resolution feature maps. In SAGAN, details can be generated using cues from all feature locations. Moreover, the discriminator can check that highly detailed features in distant portions of the image are consistent with each other. Furthermore, recent work has shown that generator conditioning affects GAN performance. Leveraging this insight, we apply spectral normalization to the GAN generator and find that this improves training dynamics. The proposed SAGAN performs better than prior work, boosting the best published Inception score from 36.8 to 52.52 and reducing Fréchet Inception distance from 27.62 to 18.65 on the challenging ImageNet dataset. Visualization of the attention layers shows that the generator leverages neighborhoods that correspond to object shapes rather than local regions of fixed shape.
****
Circuit-GNN: Graph Neural Networks for Distributed Circuit Design
Guo Zhang, Hao He, Dina Katabi
We present Circuit-GNN, a graph neural network (GNN) model for designing distributed circuits. Today, designing distributed circuits is a slow process that can take months from an expert engineer. Our model both automates and speeds up the process. The model learns to simulate the electromagnetic (EM) properties of distributed circuits. Hence, it can be used to replace traditional EM simulators, which typically take tens of minutes for each design iteration. Further, by leveraging neural networks' differentiability, we can use our model to solve the inverse problem – i.e., given desirable EM specifications, we propagate the gradient to optimize the circuit parameters and topology to satisfy the specifications.

We exploit the flexibility of GNN to create one model that works for different circuit topologies. We compare our model with a commercial simulator showing that it reduces simulation time by four orders of magnitude. We also demonstrate the value of our model by using it to design a Terahertz channelizer, a difficult task that requires a specialized expert. The results show that our model produces a channelizer whose performance is as good as a manually optimized design, and can save the expert several weeks of topology and parameter optimization. Most interestingly, our model comes up with new designs that differ from the limited templates commonly used by engineers in the field, hence significantly expanding the design space.

****

LatentGNN: Learning Efficient Non-local Relations for Visual Recognition

Songyang Zhang, Xuming He, Shipeng Yan

Capturing long-range dependencies in feature representations is crucial for many visual recognition tasks. Despite recent successes of deep convolutional networks, it remains challenging to model non-local context relations between visual features. A promising strategy is to model the feature context by a fully-connected graph neural network (GNN), which augments traditional convolutional features with an estimated non-local context representation. However, most GNN-based approaches require computing a dense graph affinity matrix and hence have difficulty in scaling up to tackle complex real-world visual problems. In this work, we propose an efficient and yet flexible non-local relation representation based on a novel class of graph neural networks. Our key idea is to introduce a latent space to reduce the complexity of graph, which allows us to use a low-rank representation for the graph affinity matrix and to achieve a linear complexity in computation. Extensive experimental evaluations on three major visual recognition tasks show that our method outperforms the prior works with a large margin while maintaining a low computation cost.

****

Neural Collaborative Subspace Clustering

Tong Zhang, Pan Ji, Mehrtash Harandi, Wenbing Huang, Hongdong Li

We introduce the Neural Collaborative Subspace Clustering, a neural model that discovers clusters of data points drawn from a union of low-dimensional subspaces. In contrast to previous attempts, our model runs without the aid of spectral clustering. This makes our algorithm one of the kinds that can gracefully scale to large datasets. At its heart, our neural model benefits from a classifier which determines whether a pair of points lies on the same subspace or not. Essential to our model is the construction of two affinity matrices, one from the classifier and the other from a notion of subspace self-expressiveness, to supervise training in a collaborative scheme. We thoroughly assess and contrast the performance of our model against various state-of-the-art clustering algorithms including deep subspace-based ones.

****

Incremental Randomized Sketching for Online Kernel Learning

Xiao Zhang, Shizhong Liao

Randomized sketching has been used in offline kernel learning, but it cannot be applied directly to online kernel learning due to the lack of incremental maintenances for randomized sketches with regret guarantees. To address these issues, we propose a novel incremental randomized sketching approach for online kernel learning, which has efficient incremental maintenances with theoretical guarantees. We construct two incremental randomized sketches using the sparse transform matrix and the sampling matrix for kernel matrix approximation, update the incremental randomized sketches using rank-1 modifications, and construct an time-varying explicit feature mapping for online kernel learning. We prove that the proposed incremental randomized sketching is statistically unbiased for the matrix product approximation, obtains a $1 + \epsilon$ relative-error bound for the kernel matrix approximation, enjoys a sublinear regret bound for online kernel learning, and has constant time and space complexities at each round for incremental maintenances. Experimental results demonstrate that the incremental randomized sketching achieves a better learning performance in terms of accuracy and effici

ency even in adversarial environments.
****
Bridging Theory and Algorithm for Domain Adaptation
Yuchen Zhang, Tianle Liu, Mingsheng Long, Michael Jordan
This paper addresses the problem of unsupervised domain adaption from theoretical and algorithmic perspectives. Existing domain adaptation theories naturally imply minimax optimization algorithms, which connect well with the domain adaptation methods based on adversarial learning. However, several disconnections still exist and form the gap between theory and algorithm. We extend previous theories (Mansour et al., 2009c; Ben-David et al., 2010) to multiclass classification in domain adaptation, where classifiers based on the scoring functions and margin loss are standard choices in algorithm design. We introduce Margin Disparity Discrepancy, a novel measurement with rigorous generalization bounds, tailored to the distribution comparison with the asymmetric margin loss, and to the minimax optimization for easier training. Our theory can be seamlessly transformed into an adversarial learning algorithm for domain adaptation, successfully bridging the gap between theory and algorithm. A series of empirical studies show that our algorithm achieves the state of the art accuracies on challenging domain adaptation tasks.
****
Adaptive Regret of Convex and Smooth Functions
Lijun Zhang, Tie-Yan Liu, Zhi-Hua Zhou
We investigate online convex optimization in changing environments, and choose the adaptive regret as the performance measure. The goal is to achieve a small regret over every interval so that the comparator is allowed to change over time. Different from previous works that only utilize the convexity condition, this paper further exploits smoothness to improve the adaptive regret. To this end, we develop novel adaptive algorithms for convex and smooth functions, and establish problem-dependent regret bounds over any interval. Our regret bounds are comparable to existing results in the worst case, and become much tighter when the comparator has a small loss.
****
Random Function Priors for Correlation Modeling
Aonan Zhang, John Paisley
The likelihood model of high dimensional data X_n can often be expressed as $p(X_n|Z_n,\theta)$, where $\theta\mathrel{\mathop:}=(\theta_k)_{k\in[K]}$ is a collection of hidden features shared across objects, indexed by n, and Z_n is a non-negative factor loading vector with K entries where Z_{nk} indicates the strength of θ_k used to express X_n. In this paper, we introduce random function priors for Z_n for modeling correlations among its K dimensions Z_{n1} through Z_{nK}, which we call population random measure embedding (PRME). Our model can be viewed as a generalized paintbox model \cite{Broderick13} using random functions, and can be learned efficiently with neural networks via amortized variational inference. We derive our Bayesian nonparametric method by applying a representation theorem on separately exchangeable discrete random measures.
****
Co-Representation Network for Generalized Zero-Shot Learning
Fei Zhang, Guangming Shi
Generalized zero-shot learning is a significant topic but faced with bias problem, which leads to unseen classes being easily misclassified into seen classes. Hence we propose a embedding model called co-representation network to learn a more uniform visual embedding space that effectively alleviates the bias problem and helps with classification. We mathematically analyze our model and find it learns a projection with high local linearity, which is proved to cause less bias problem. The network consists of a cooperation module for representation and a relation module for classification, it is simple in structure and can be easily trained in an end-to-end manner. Experiments show that our method outperforms existing generalized zero-shot learning methods on several benchmark datasets.
****

SOLAR: Deep Structured Representations for Model-Based Reinforcement Learning
Marvin Zhang, Sharad Vikram, Laura Smith, Pieter Abbeel, Matthew Johnson, Sergey Levine
Model-based reinforcement learning (RL) has proven to be a data efficient approach for learning control tasks but is difficult to utilize in domains with complex observations such as images. In this paper, we present a method for learning representations that are suitable for iterative model-based policy improvement, even when the underlying dynamical system has complex dynamics and image observations, in that these representations are optimized for inferring simple dynamics and cost models given data from the current policy. This enables a model-based RL method based on the linear-quadratic regulator (LQR) to be used for systems with image observations. We evaluate our approach on a range of robotics tasks, including manipulation with a real-world robotic arm directly from images. We find that our method produces substantially better final performance than other model-based RL methods while being significantly more efficient than model-free RL.
****
A Composite Randomized Incremental Gradient Method
Junyu Zhang, Lin Xiao
We consider the problem of minimizing the composition of a smooth function (which can be nonconvex) and a smooth vector mapping, where both of them can be express as the average of a large number of components. We propose a composite randomized incremental gradient method by extending the SAGA framework. The gradient sample complexity of our method matches that of several recently developed methods based on SVRG in the general case. However, for structured problems where linear convergence rates can be obtained, our method can be much better for ill-conditioned problems. In addition, when the finite-sum structure only appear for the inner mapping, the sample complexity of our method is the same as that of SAGA for minimizing finite sum of smooth nonconvex functions, despite the additional outer composition and the stochastic composite gradients being biased in our case.
****
Fast and Stable Maximum Likelihood Estimation for Incomplete Multinomial Models
Chenyang Zhang, Guosheng Yin
We propose a fixed-point iteration approach to the maximum likelihood estimation for the incomplete multinomial model, which provides a unified framework for ranking data analysis. Incomplete observations typically fall in a subset of categories, and thus cannot be distinguished as belonging to a unique category. We develop a minorization–maximization (MM) type of algorithm, which requires relatively fewer iterations and shorter time to achieve convergence. Under such a general framework, incomplete multinomial models can be reformulated to include several well-known ranking models as special cases, such as the Bradley–Terry, Plackett–Luce models and their variants. The simple form of iteratively updating equations in our algorithm involves only basic matrix operations, which makes it efficient and easy to implement with large data. Experimental results show that our algorithm runs faster than existing methods on synthetic data and real data.
****
Theoretically Principled Trade-off between Robustness and Accuracy
Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, Michael Jordan
We identify a trade-off between robustness and accuracy that serves as a guiding principle in the design of defenses against adversarial examples. Although this problem has been widely studied empirically, much remains unknown concerning the theory underlying this trade-off. In this work, we decompose the prediction error for adversarial examples (robust error) as the sum of the natural (classification) error and boundary error, and provide a differentiable upper bound using the theory of classification-calibrated loss, which is shown to be the tightest possible upper bound uniform over all probability distributions and measurable predictors. Inspired by our theoretical analysis, we also design a new defense method, TRADES, to trade adversarial robustness off against accuracy. Our proposed algorithm performs well experimentally in real-world datasets. The methodology

is the foundation of our entry to the NeurIPS 2018 Adversarial Vision Challenge in which we won the 1st place out of 2,000 submissions, surpassing the runner-up approach by 11.41% in terms of mean L_2 perturbation distance.
****

Learning Novel Policies For Tasks
Yunbo Zhang, Wenhao Yu, Greg Turk
In this work, we present a reinforcement learning algorithm that can find a variety of policies (novel policies) for a task that is given by a task reward function. Our method does this by creating a second reward function that recognizes previously seen state sequences and rewards those by novelty, which is measured using autoencoders that have been trained on state sequences from previously discovered policies. We present a two-objective update technique for policy gradient algorithms in which each update of the policy is a compromise between improving the task reward and improving the novelty reward. Using this method, we end up with a collection of policies that solves a given task as well as carrying out action sequences that are distinct from one another. We demonstrate this method on maze navigation tasks, a reaching task for a simulated robot arm, and a locomotion task for a hopper. We also demonstrate the effectiveness of our approach on deceptive tasks in which policy gradient methods often get stuck.
****

Greedy Orthogonal Pivoting Algorithm for Non-Negative Matrix Factorization
Kai Zhang, Sheng Zhang, Jun Liu, Jun Wang, Jie Zhang
Non-negative matrix factorization is a powerful tool for learning useful representations in the data and has been widely applied in many problems such as data mining and signal processing. Orthogonal NMF, which can improve the locality of decomposition, has drawn considerable interest in solving clustering problems in recent years. However, imposing simultaneous non-negative and orthogonal structure can be quite difficult, and so existing algorithms can only solve it approximately. To address this challenge, we propose an innovative procedure called Greedy Orthogonal Pivoting Algorithm (GOPA). The GOPA algorithm fully exploits the sparsity of non-negative orthogonal solutions to break the global problem into a series of local optimizations, in which an adaptive subset of coordinates are updated in a greedy, closed-form manner. The biggest advantage of GOPA is that it promotes exact orthogonality and provides solid empirical evidence that stronger orthogonality does contribute favorably to better clustering performance. On the other hand, we further design randomized and parallel version of GOPA, which can further reduce the computational cost and improve accuracy, making it suitable for large data.
****

Interpreting Adversarially Trained Convolutional Neural Networks
Tianyuan Zhang, Zhanxing Zhu
We attempt to interpret how adversarially trained convolutional neural networks (AT-CNNs) recognize objects. We design systematic approaches to interpret AT-CNNs in both qualitative and quantitative ways and compare them with normally trained models. Surprisingly, we find that adversarial training alleviates the texture bias of standard CNNs when trained on object recognition tasks, and helps CNNs learn a more shape-biased representation. We validate our hypothesis from two aspects. First, we compare the salience maps of AT-CNNs and standard CNNs on clean images and images under different transformations. The comparison could visually show that the prediction of the two types of CNNs is sensitive to dramatically different types of features. Second, to achieve quantitative verification, we construct additional test datasets that destroy either textures or shapes, such as style-transferred version of clean data, saturated images and patch-shuffled ones, and then evaluate the classification accuracy of AT-CNNs and normal CNNs on these datasets. Our findings shed some light on why AT-CNNs are more robust than those normally trained ones and contribute to a better understanding of adversarial training over CNNs from an interpretation perspective.
****

Adaptive Monte Carlo Multiple Testing via Multi-Armed Bandits
Martin Zhang, James Zou, David Tse

Monte Carlo (MC) permutation test is considered the gold standard for statistica
l hypothesis testing, especially when standard parametric assumptions are not cl
ear or likely to fail. However, in modern data science settings where a large nu
mber of hypothesis tests need to be performed simultaneously, it is rarely used
due to its prohibitive computational cost. In genome-wide association studies, f
or example, the number of hypothesis tests m is around 10^6 while the number
 of MC samples n for each test could be greater than 10^8, totaling more tha
n nm=10^{14} samples. In this paper, we propose \texttt{A}daptive \texttt{M}
C multiple \texttt{T}esting (\texttt{AMT}) to estimate MC p-values and control f
alse discovery rate in multiple testing. The algorithm outputs the same result a
s the standard full MC approach with high probability while requiring only $\til
de{O}(\sqrt{n}m)$ samples. This sample complexity is shown to be optimal. On a P
arkinson GWAS dataset, the algorithm reduces the running time from 2 months for
full MC to an hour. The \texttt{AMT} algorithm is derived based on the theory of
 multi-armed bandits.
****
On Learning Invariant Representations for Domain Adaptation
Han Zhao, Remi Tachet Des Combes, Kun Zhang, Geoffrey Gordon
Due to the ability of deep neural nets to learn rich representations, recent adv
ances in unsupervised domain adaptation have focused on learning domain-invarian
t features that achieve a small error on the source domain. The hope is that the
 learnt representation, together with the hypothesis learnt from the source doma
in, can generalize to the target domain. In this paper, we first construct a sim
ple counterexample showing that, contrary to common belief, the above conditions
 are not sufficient to guarantee successful domain adaptation. In particular, th
e counterexample exhibits conditional shift: the class-conditional distributions
 of input features change between source and target domains. To give a sufficien
t condition for domain adaptation, we propose a natural and interpretable genera
lization upper bound that explicitly takes into account the aforementioned shift
. Moreover, we shed new light on the problem by proving an information-theoretic
 lower bound on the joint error of any domain adaptation method that attempts to
 learn invariant representations. Our result characterizes a fundamental tradeof
f between learning invariant representations and achieving small joint error on
both domains when the marginal label distributions differ from source to target.
 Finally, we conduct experiments on real-world datasets that corroborate our the
oretical findings. We believe these insights are helpful in guiding the future d
esign of domain adaptation and representation learning algorithms.
****
Metric-Optimized Example Weights
Sen Zhao, Mahdi Milani Fard, Harikrishna Narasimhan, Maya Gupta
Real-world machine learning applications often have complex test metrics, and ma
y have training and test data that are not identically distributed. Motivated by
 known connections between complex test metrics and cost-weighted learning, we p
ropose addressing these issues by using a weighted loss function with a standard
 loss, where the weights on the training examples are learned to optimize the te
st metric on a validation set. These metric-optimized example weights can be lea
rned for any test metric, including black box and customized ones for specific a
pplications. We illustrate the performance of the proposed method on diverse pub
lic benchmark datasets and real-world applications. We also provide a generaliza
tion bound for the method.
****
Improving Neural Network Quantization without Retraining using Outlier Channel S
plitting
Ritchie Zhao, Yuwei Hu, Jordan Dotzel, Chris De Sa, Zhiru Zhang
Quantization can improve the execution latency and energy efficiency of neural n
etworks on both commodity GPUs and specialized accelerators. The majority of exi
sting literature focuses on training quantized DNNs, while this work examines th
e less-studied topic of quantizing a floating-point model without (re)training.
DNN weights and activations follow a bell-shaped distribution post-training, whi
le practical hardware uses a linear quantization grid. This leads to challenges

in dealing with outliers in the distribution. Prior work has addressed this by clipping the outliers or using specialized hardware. In this work, we propose outlier channel splitting (OCS), which duplicates channels containing outliers, then halves the channel values. The network remains functionally identical, but affected outliers are moved toward the center of the distribution. OCS requires no additional training and works on commodity hardware. Experimental evaluation on ImageNet classification and language modeling shows that OCS can outperform state-of-the-art clipping techniques with only minor overhead.
****
Maximum Entropy-Regularized Multi-Goal Reinforcement Learning
Rui Zhao, Xudong Sun, Volker Tresp
In Multi-Goal Reinforcement Learning, an agent learns to achieve multiple goals with a goal-conditioned policy. During learning, the agent first collects the trajectories into a replay buffer, and later these trajectories are selected randomly for replay. However, the achieved goals in the replay buffer are often biased towards the behavior policies. From a Bayesian perspective, when there is no prior knowledge about the target goal distribution, the agent should learn uniformly from diverse achieved goals. Therefore, we first propose a novel multi-goal RL objective based on weighted entropy. This objective encourages the agent to maximize the expected return, as well as to achieve more diverse goals. Secondly, we developed a maximum entropy-based prioritization framework to optimize the proposed objective. For evaluation of this framework, we combine it with Deep Deterministic Policy Gradient, both with or without Hindsight Experience Replay. On a set of multi-goal robotic tasks of OpenAI Gym, we compare our method with other baselines and show promising improvements in both performance and sample-efficiency.
****
Stochastic Iterative Hard Thresholding for Graph-structured Sparsity Optimization
Baojian Zhou, Feng Chen, Yiming Ying
Stochastic optimization algorithms update models with cheap per-iteration costs sequentially, which makes them amenable for large-scale data analysis. Such algorithms have been widely studied for structured sparse models where the sparsity information is very specific, e.g., convex sparsity-inducing norms or ℓ^0-norm. However, these norms cannot be directly applied to the problem of complex (non-convex) graph-structured sparsity models, which have important application in disease outbreak and social networks, etc. In this paper, we propose a stochastic gradient-based method for solving graph-structured sparsity constraint problems, not restricted to the least square loss. We prove that our algorithm enjoys a linear convergence up to a constant error, which is competitive with the counterparts in the batch learning setting. We conduct extensive experiments to show the efficiency and effectiveness of the proposed algorithms.
****
Lower Bounds for Smooth Nonconvex Finite-Sum Optimization
Dongruo Zhou, Quanquan Gu
Smooth finite-sum optimization has been widely studied in both convex and nonconvex settings. However, existing lower bounds for finite-sum optimization are mostly limited to the setting where each component function is (strongly) convex, while the lower bounds for nonconvex finite-sum optimization remain largely unsolved. In this paper, we study the lower bounds for smooth nonconvex finite-sum optimization, where the objective function is the average of n nonconvex component functions. We prove tight lower bounds for the complexity of finding ϵ-suboptimal point and ϵ-approximate stationary point in different settings, for a wide regime of the smallest eigenvalue of the Hessian of the objective function (or each component function). Given our lower bounds, we can show that existing algorithms including {KatyushaX} \citep{allen2018katyushax}, {Natasha} \citep{allen2017natasha} and {StagewiseKatyusha} \citep{yang2018does} have achieved optimal {Incremental First-order Oracle} (IFO) complexity (i.e., number of IFO calls) up to logarithm factors for nonconvex finite-sum optimization. We also point out potential ways to further improve these complexity results, in te

rms of making stronger assumptions or by a different convergence analysis.
****
Lipschitz Generative Adversarial Nets
Zhiming Zhou, Jiadong Liang, Yuxuan Song, Lantao Yu, Hongwei Wang, Weinan Zhang,
 Yong Yu, Zhihua Zhang
In this paper we show that generative adversarial networks (GANs) without restri
ction on the discriminative function space commonly suffer from the problem that
 the gradient produced by the discriminator is uninformative to guide the genera
tor. By contrast, Wasserstein GAN (WGAN), where the discriminative function is r
estricted to 1-Lipschitz, does not suffer from such a gradient uninformativeness
 problem. We further show in the paper that the model with a compact dual form o
f Wasserstein distance, where the Lipschitz condition is relaxed, may also theor
etically suffer from this issue. This implies the importance of Lipschitz condit
ion and motivates us to study the general formulation of GANs with Lipschitz con
straint, which leads to a new family of GANs that we call Lipschitz GANs (LGANs)
. We show that LGANs guarantee the existence and uniqueness of the optimal discr
iminative function as well as the existence of a unique Nash equilibrium. We pro
ve that LGANs are generally capable of eliminating the gradient uninformativenes
s problem. According to our empirical analysis, LGANs are more stable and genera
te consistently higher quality samples compared with WGAN.
****
Toward Understanding the Importance of Noise in Training Neural Networks
Mo Zhou, Tianyi Liu, Yan Li, Dachao Lin, Enlu Zhou, Tuo Zhao
Numerous empirical evidence has corroborated that the noise plays a crucial rule
 in effective and efficient training of deep neural networks. The theory behind,
 however, is still largely unknown. This paper studies this fundamental problem
through training a simple two-layer convolutional neural network model. Although
 training such a network requires to solve a non-convex optimization problem wit
h a spurious local optimum and a global optimum, we prove that a perturbed gradi
ent descent algorithm in conjunction with noise annealing is guaranteed to conve
rge to a global optimum in polynomial time with arbitrary initialization. This i
mplies that the noise enables the algorithm to efficiently escape from the spuri
ous local optimum. Numerical experiments are provided to support our theory.
****
BayesNAS: A Bayesian Approach for Neural Architecture Search
Hongpeng Zhou, Minghao Yang, Jun Wang, Wei Pan
One-Shot Neural Architecture Search (NAS) is a promising method to significantly
 reduce search time without any separate training. It can be treated as a Networ
k Compression problem on the architecture parameters from an over-parameterized
network. However, there are two issues associated with most one-shot NAS methods
. First, dependencies between a node and its predecessors and successors are oft
en disregarded which result in improper treatment over zero operations. Second,
architecture parameters pruning based on their magnitude is questionable. In thi
s paper, we employ the classic Bayesian learning approach to alleviate these two
 issues by modeling architecture parameters using hierarchical automatic relevan
ce determination (HARD) priors. Unlike other NAS methods, we train the over-para
meterized network for only one epoch then update the architecture. Impressively,
 this enabled us to find the architecture in both proxy and proxyless tasks on C
IFAR-10 within only 0.2 GPU days using a single GPU. As a byproduct, our approac
h can be transferred directly to compress convolutional neural networks by enfor
cing structural sparsity which achieves extremely sparse networks without accura
cy deterioration.
****
Transferable Clean-Label Poisoning Attacks on Deep Neural Nets
Chen Zhu, W. Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, Tom Goldst
ein
In this paper, we explore clean-label poisoning attacks on deep convolutional ne
tworks with access to neither the network’s output nor its architecture or param
eters. Our goal is to ensure that after injecting the poisons into the training
data, a model with unknown architecture and parameters trained on that data will

misclassify the target image into a specific class. To achieve this goal, we generate multiple poison images from the base class by adding small perturbations which cause the poison images to trap the target image within their convex polytope in feature space. We also demonstrate that using Dropout during crafting of the poisons and enforcing this objective in multiple layers enhances transferability, enabling attacks against both the transfer learning and end-to-end training settings. We demonstrate transferable attack success rates of over 50% by poisoning only 1% of the training set.
****

Improved Dynamic Graph Learning through Fault-Tolerant Sparsification

Chunjiang Zhu, Sabine Storandt, Kam-Yiu Lam, Song Han, Jinbo Bi

Graph sparsification has been used to improve the computational cost of learning over graphs, e.g., Laplacian-regularized estimation and graph semi-supervised learning (SSL). However, when graphs vary over time, repeated sparsification requires polynomial order computational cost per update. We propose a new type of graph sparsification namely fault-tolerant (FT) sparsification to significantly reduce the cost to only a constant. Then the computational cost of subsequent graph learning tasks can be significantly improved with limited loss in their accuracy. In particular, we give theoretical analyze to upper bound the loss in the accuracy of the subsequent Laplacian-regularized estimation and graph SSL, due to the FT sparsification. In addition, FT spectral sparsification can be generalized to FT cut sparsification, for cut-based graph learning. Extensive experiments have confirmed the computational efficiencies and accuracies of the proposed methods for learning on dynamic graphs.
****

Poission Subsampled Rényi Differential Privacy

Yuqing Zhu, Yu-Xiang Wang

We consider the problem of privacy-amplification by under the Renyi Differential Privacy framework. This is the main technique underlying the moments accountants (Abadi et al., 2016) for differentially private deep learning. Unlike previous attempts on this problem which deals with Sampling with Replacement, we consider the Poisson subsampling scheme which selects each data point independently with a coin toss. This allows us to significantly simplify and tighten the bounds for the RDP of subsampled mechanisms and derive numerically stable approximation schemes. In particular, for subsampled Gaussian mechanism and subsampled Laplace mechanism, we prove an analytical formula of their RDP that exactly matches the lower bound. The result is the first of its kind and we numerically demonstrate an order of magnitude improvement in the privacy-utility tradeoff.
****

Learning Classifiers for Target Domain with Limited or No Labels

Pengkai Zhu, Hanxiao Wang, Venkatesh Saligrama

In computer vision applications, such as domain adaptation (DA), few shot learning (FSL) and zero-shot learning (ZSL), we encounter new objects and environments, for which insufficient examples exist to allow for training “models from scratch,” and methods that adapt existing models, trained on the presented training environment, to the new scenario are required. We propose a novel visual attribute encoding method that encodes each image as a low-dimensional probability vector composed of prototypical part-type probabilities. The prototypes are learnt to be representative of all training data. At test-time we utilize this encoding as an input to a classifier. At test-time we freeze the encoder and only learn/adapt the classifier component to limited annotated labels in FSL; new semantic attributes in ZSL. We conduct extensive experiments on benchmark datasets. Our method outperforms state-of-art methods trained for the specific contexts (ZSL, FSL, DA).
****

The Anisotropic Noise in Stochastic Gradient Descent: Its Behavior of Escaping from Sharp Minima and Regularization Effects

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, Jinwen Ma

Understanding the behavior of stochastic gradient descent (SGD) in the context of deep neural networks has raised lots of concerns recently. Along this line, we

study a general form of gradient based optimization dynamics with unbiased noise, which unifies SGD and standard Langevin dynamics. Through investigating this general optimization dynamics, we analyze the behavior of SGD on escaping from minima and its regularization effects. A novel indicator is derived to characterize the efficiency of escaping from minima through measuring the alignment of noise covariance and the curvature of loss function. Based on this indicator, two conditions are established to show which type of noise structure is superior to isotropic noise in term of escaping efficiency. We further show that the anisotropic noise in SGD satisfies the two conditions, and thus helps to escape from sharp and poor minima effectively, towards more stable and flat minima that typically generalize well. We systematically design various experiments to verify the benefits of the anisotropic noise, compared with full gradient descent plus isotropic diffusion (i.e. Langevin dynamics).
****
Surrogate Losses for Online Learning of Stepsizes in Stochastic Non-Convex Optimization
Zhenxun Zhuang, Ashok Cutkosky, Francesco Orabona
Stochastic Gradient Descent (SGD) has played a central role in machine learning. However, it requires a carefully hand-picked stepsize for fast convergence, which is notoriously tedious and time-consuming to tune. Over the last several years, a plethora of adaptive gradient-based algorithms have emerged to ameliorate this problem. In this paper, we propose new surrogate losses to cast the problem of learning the optimal stepsizes for the stochastic optimization of a non-convex smooth objective function onto an online convex optimization problem. This allows the use of no-regret online algorithms to compute optimal stepsizes on the fly. In turn, this results in a SGD algorithm with self-tuned stepsizes that guarantees convergence rates that are automatically adaptive to the level of noise.
****
Latent Normalizing Flows for Discrete Sequences
Zachary Ziegler, Alexander Rush
Normalizing flows are a powerful class of generative models for continuous random variables, showing both strong model flexibility and the potential for non-autoregressive generation. These benefits are also desired when modeling discrete random variables such as text, but directly applying normalizing flows to discrete sequences poses significant additional challenges. We propose a VAE-based generative model which jointly learns a normalizing flow-based distribution in the latent space and a stochastic mapping to an observed discrete space. In this setting, we find that it is crucial for the flow-based distribution to be highly multimodal. To capture this property, we propose several normalizing flow architectures to maximize model flexibility. Experiments consider common discrete sequence tasks of character-level language modeling and polyphonic music generation. Our results indicate that an autoregressive flow-based model can match the performance of a comparable autoregressive baseline, and a non-autoregressive flow-based model can improve generation speed with a penalty to performance.
****
Beating Stochastic and Adversarial Semi-bandits Optimally and Simultaneously
Julian Zimmert, Haipeng Luo, Chen-Yu Wei
We develop the first general semi-bandit algorithm that simultaneously achieves $\mathcal{O}(\log T)$ regret for stochastic environments and $\mathcal{O}(\sqrt{T})$ regret for adversarial environments without knowledge of the regime or the number of rounds T. The leading problem-dependent constants of our bounds are not only optimal in some worst-case sense studied previously, but also optimal for two concrete instances of semi-bandit problems. Our algorithm and analysis extend the recent work of (Zimmert & Seldin, 2019) for the special case of multi-armed bandits, but importantly requires a novel hybrid regularizer designed specifically for semi-bandit. Experimental results on synthetic data show that our algorithm indeed performs well uniformly over different environments. We finally provide a preliminary extension of our results to the full bandit feedback.
****
Fast Context Adaptation via Meta-Learning

Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, Shimon Whiteson

We propose CAVIA for meta-learning, a simple extension to MAML that is less prone to meta-overfitting, easier to parallelise, and more interpretable. CAVIA partitions the model parameters into two parts: context parameters that serve as additional input to the model and are adapted on individual tasks, and shared parameters that are meta-trained and shared across tasks. At test time, only the context parameters are updated, leading to a low-dimensional task representation. We show empirically that CAVIA outperforms MAML for regression, classification, and reinforcement learning. Our experiments also highlight weaknesses in current benchmarks, in that the amount of adaptation needed in some cases is small.

****

Natural Analysts in Adaptive Data Analysis

Tijana Zrnic, Moritz Hardt

Adaptive data analysis is frequently criticized for its pessimistic generalization guarantees. The source of these pessimistic bounds is a model that permits arbitrary, possibly adversarial analysts that optimally use information to bias results. While being a central issue in the field, still lacking are notions of natural analysts that allow for more optimistic bounds faithful to the reality that typical analysts aren’t adversarial. In this work, we propose notions of natural analysts that smoothly interpolate between the optimal non-adaptive bounds and the best-known adaptive generalization bounds. To accomplish this, we model the analyst’s knowledge as evolving according to the rules of an unknown dynamical system that takes in revealed information and outputs new statistical queries to the data. This allows us to restrict the analyst through different natural control-theoretic notions. One such notion corresponds to a recency bias, formalizing an inability to arbitrarily use distant information. Another complementary notion formalizes an anchoring bias, a tendency to weight initial information more strongly. Both notions come with quantitative parameters that smoothly interpolate between the non-adaptive case and the fully adaptive case, allowing for a rich spectrum of intermediate analysts that are neither non-adaptive nor adversarial. Natural not only from a cognitive perspective, we show that our notions also capture standard optimization methods, like gradient descent in various settings. This gives a new interpretation to the fact that gradient descent tends to overfit much less than its adaptive nature might suggest.

****