Sign in to GitHub · GitHub
****
A New Representation of Successor Features for Transfer across Dissimilar Environments
Majid Abdolshah, Hung Le, Thommen Karimpanal George, Sunil Gupta, Santu Rana, Svetha Venkatesh
Transfer in reinforcement learning is usually achieved through generalisation across tasks. Whilst many studies have investigated transferring knowledge when the reward function changes, they have assumed that the dynamics of the environments remain consistent. Many real-world RL problems require transfer among environments with different dynamics. To address this problem, we propose an approach based on successor features in which we model successor feature functions with Gaussian Processes permitting the source successor features to be treated as noisy measurements of the target successor feature function. Our theoretical analysis proves the convergence of this approach as well as the bounded error on modelling successor feature functions with Gaussian Processes in environments with both different dynamics and rewards. We demonstrate our method on benchmark datasets and show that it outperforms current baselines.
****
Massively Parallel and Asynchronous Tsetlin Machine Architecture Supporting Almost Constant-Time Scaling
Kuruge Darshana Abeyrathna, Bimal Bhattarai, Morten Goodwin, Saeed Rahimi Gorji, Ole-Christoffer Granmo, Lei Jiao, Rupsa Saha, Rohan K. Yadav
Using logical clauses to represent patterns, Tsetlin Machine (TM) have recently obtained competitive performance in terms of accuracy, memory footprint, energy, and learning speed on several benchmarks. Each TM clause votes for or against a particular class, with classification resolved using a majority vote. While the evaluation of clauses is fast, being based on binary operators, the voting makes it necessary to synchronize the clause evaluation, impeding parallelization. In this paper, we propose a novel scheme for desynchronizing the evaluation of clauses, eliminating the voting bottleneck. In brief, every clause runs in its own thread for massive native parallelism. For each training example, we keep track of the class votes obtained from the clauses in local voting tallies. The local voting tallies allow us to detach the processing of each clause from the rest of the clauses, supporting decentralized learning. This means that the TM most of the time will operate on outdated voting tallies. We evaluated the proposed parallelization across diverse learning tasks and it turns out that our decentralized TM learning algorithm copes well with working on outdated data, resulting in no significant loss in learning accuracy. Furthermore, we show that the approach provides up to 50 times faster learning. Finally, learning time is almost constant for reasonable clause amounts (employing from 20 to 7,000 clauses on a Tesla V100 GPU). For sufficiently large clause numbers, computation time increases approximately proportionally. Our parallel and asynchronous architecture thus allows processing of more massive datasets and operating with more clauses for higher accuracy.
****
Debiasing Model Updates for Improving Personalized Federated Training
Durmus Alp Emre Acar, Yue Zhao, Ruizhao Zhu, Ramon Matas, Matthew Mattina, Paul Whatmough, Venkatesh Saligrama
We propose a novel method for federated learning that is customized specifically to the objective of a given edge device. In our proposed method, a server trains a global meta-model by collaborating with devices without actually sharing data. The trained global meta-model is then personalized locally by each device to meet its specific objective. Different from the conventional federated learning setting, training customized models for each device is hindered by both the inherent data biases of the various devices, as well as the requirements imposed by the federated architecture. We propose gradient correction methods leveraging prior works, and explicitly de-bias the meta-model in the distributed heterogeneous data setting to learn personalized device models. We present convergence guarantees of our method for strongly convex, convex and nonconvex meta objectives. W

e empirically evaluate the performance of our method on benchmark datasets and d emonstrate significant communication savings.
****

Memory Efficient Online Meta Learning
Durmus Alp Emre Acar, Ruizhao Zhu, Venkatesh Saligrama
We propose a novel algorithm for online meta learning where task instances are s equentially revealed with limited supervision and a learner is expected to meta learn them in each round, so as to allow the learner to customize a task-specifi c model rapidly with little task-level supervision. A fundamental concern arisin g in online meta-learning is the scalability of memory as more tasks are viewed over time. Heretofore, prior works have allowed for perfect recall leading to li near increase in memory with time. Different from prior works, in our method, pr ior task instances are allowed to be deleted. We propose to leverage prior task instances by means of a fixed-size state-vector, which is updated sequentially. Our theoretical analysis demonstrates that our proposed memory efficient online learning (MOML) method suffers sub-linear regret with convex loss functions and sub-linear local regret for nonconvex losses. On benchmark datasets we show that our method can outperform prior works even though they allow for perfect recall .
****

Robust Testing and Estimation under Manipulation Attacks
Jayadev Acharya, Ziteng Sun, Huanyu Zhang
We study robust testing and estimation of discrete distributions in the strong c ontamination model. Our results cover both centralized setting and distributed s etting with general local information constraints including communication and LD P constraints. Our technique relates the strength of manipulation attacks to the earth-mover distance using Hamming distance as the metric between messages (sam ples) from the users. In the centralized setting, we provide optimal error bound s for both learning and testing. Our lower bounds under local information constr aints build on the recent lower bound methods in distributed inference. In the c ommunication constrained setting, we develop novel algorithms based on random ha shing and an L1-L1 isometry.
****

GP-Tree: A Gaussian Process Classifier for Few-Shot Incremental Learning
Idan Achituve, Aviv Navon, Yochai Yemini, Gal Chechik, Ethan Fetaya
Gaussian processes (GPs) are non-parametric, flexible, models that work well in many tasks. Combining GPs with deep learning methods via deep kernel learning (D KL) is especially compelling due to the strong representational power induced by the network. However, inference in GPs, whether with or without DKL, can be com putationally challenging on large datasets. Here, we propose GP-Tree, a novel me thod for multi-class classification with Gaussian processes and DKL. We develop a tree-based hierarchical model in which each internal node of the tree fits a G P to the data using the P{ó}lya-Gamma augmentation scheme. As a result, our meth od scales well with both the number of classes and data size. We demonstrate the effectiveness of our method against other Gaussian process training baselines, and we show how our general GP approach achieves improved accuracy on standard i ncremental few-shot learning benchmarks.
****

f-Domain Adversarial Learning: Theory and Algorithms
David Acuna, Guojun Zhang, Marc T. Law, Sanja Fidler
Unsupervised domain adaptation is used in many machine learning applications whe re, during training, a model has access to unlabeled data in the target domain, and a related labeled dataset. In this paper, we introduce a novel and general d omain-adversarial framework. Specifically, we derive a novel generalization boun d for domain adaptation that exploits a new measure of discrepancy between distr ibutions based on a variational characterization of f-divergences. It recovers t he theoretical results from Ben-David et al. (2010a) as a special case and suppo rts divergences used in practice. Based on this bound, we derive a new algorithm ic framework that introduces a key correction in the original adversarial traini ng method of Ganin et al. (2016). We show that many regularizers and ad-hoc obje

ctives introduced over the last years in this framework are then not required to achieve performance comparable to (if not better than) state-of-the-art domain-adversarial methods. Experimental analysis conducted on real-world natural language and computer vision datasets show that our framework outperforms existing baselines, and obtains the best results for f-divergences that were not considered previously in domain-adversarial learning.
****
Towards Rigorous Interpretations: a Formalisation of Feature Attribution
Darius Afchar, Vincent Guigue, Romain Hennequin
Feature attribution is often loosely presented as the process of selecting a subset of relevant features as a rationale of a prediction. Task-dependent by nature, precise definitions of "relevance" encountered in the literature are however not always consistent. This lack of clarity stems from the fact that we usually do not have access to any notion of ground-truth attribution and from a more general debate on what good interpretations are. In this paper we propose to formalise feature selection/attribution based on the concept of relaxed functional dependence. In particular, we extend our notions to the instance-wise setting and derive necessary properties for candidate selection solutions, while leaving room for task-dependence. By computing ground-truth attributions on synthetic datasets, we evaluate many state-of-the-art attribution methods and show that, even when optimised, some fail to verify the proposed properties and provide wrong solutions.
****
Acceleration via Fractal Learning Rate Schedules
Naman Agarwal, Surbhi Goel, Cyril Zhang
In practical applications of iterative first-order optimization, the learning rate schedule remains notoriously difficult to understand and expensive to tune. We demonstrate the presence of these subtleties even in the innocuous case when the objective is a convex quadratic. We reinterpret an iterative algorithm from the numerical analysis literature as what we call the Chebyshev learning rate schedule for accelerating vanilla gradient descent, and show that the problem of mitigating instability leads to a fractal ordering of step sizes. We provide some experiments to challenge conventional beliefs about stable learning rates in deep learning: the fractal schedule enables training to converge with locally unstable updates which make negative progress on the objective.
****
A Regret Minimization Approach to Iterative Learning Control
Naman Agarwal, Elad Hazan, Anirudha Majumdar, Karan Singh
We consider the setting of iterative learning control, or model-based policy learning in the presence of uncertain, time-varying dynamics. In this setting, we propose a new performance metric, planning regret, which replaces the standard stochastic uncertainty assumptions with worst case regret. Based on recent advances in non-stochastic control, we design a new iterative algorithm for minimizing planning regret that is more robust to model mismatch and uncertainty. We provide theoretical and empirical evidence that the proposed algorithm outperforms existing methods on several benchmarks.
****
Towards the Unification and Robustness of Perturbation and Gradient Based Explanations
Sushant Agarwal, Shahin Jabbari, Chirag Agarwal, Sohini Upadhyay, Steven Wu, Himabindu Lakkaraju
As machine learning black boxes are increasingly being deployed in critical domains such as healthcare and criminal justice, there has been a growing emphasis on developing techniques for explaining these black boxes in a post hoc manner. In this work, we analyze two popular post hoc interpretation techniques: SmoothGrad which is a gradient based method, and a variant of LIME which is a perturbation based method. More specifically, we derive explicit closed form expressions for the explanations output by these two methods and show that they both converge to the same explanation in expectation, i.e., when the number of perturbed samples used by these methods is large. We then leverage this connection to establis

h other desirable properties, such as robustness, for these techniques. We also derive finite sample complexity bounds for the number of perturbations required for these methods to converge to their expected explanation. Finally, we empirically validate our theory using extensive experimentation on both synthetic and real-world datasets.
****
Label Inference Attacks from Log-loss Scores
Abhinav Aggarwal, Shiva Kasiviswanathan, Zekun Xu, Oluwaseyi Feyisetan, Nathanael Teissier
Log-loss (also known as cross-entropy loss) metric is ubiquitously used across machine learning applications to assess the performance of classification algorithms. In this paper, we investigate the problem of inferring the labels of a dataset from single (or multiple) log-loss score(s), without any other access to the dataset. Surprisingly, we show that for any finite number of label classes, it is possible to accurately infer the labels of the dataset from the reported log-loss score of a single carefully constructed prediction vector if we allow arbitrary precision arithmetic. Additionally, we present label inference algorithms (attacks) that succeed even under addition of noise to the log-loss scores and under limited precision arithmetic. All our algorithms rely on ideas from number theory and combinatorics and require no model training. We run experimental simulations on some real datasets to demonstrate the ease of running these attacks in practice.
****
Deep kernel processes
Laurence Aitchison, Adam Yang, Sebastian W. Ober
We define deep kernel processes in which positive definite Gram matrices are progressively transformed by nonlinear kernel functions and by sampling from (inverse) Wishart distributions. Remarkably, we find that deep Gaussian processes (DGPs), Bayesian neural networks (BNNs), infinite BNNs, and infinite BNNs with bottlenecks can all be written as deep kernel processes. For DGPs the equivalence arises because the Gram matrix formed by the inner product of features is Wishart distributed, and as we show, standard isotropic kernels can be written entirely in terms of this Gram matrix — we do not need knowledge of the underlying features. We define a tractable deep kernel process, the deep inverse Wishart process, and give a doubly-stochastic inducing-point variational inference scheme that operates on the Gram matrices, not on the features, as in DGPs. We show that the deep inverse Wishart process gives superior performance to DGPs and infinite BNNs on fully-connected baselines.
****
How Does Loss Function Affect Generalization Performance of Deep Learning? Application to Human Age Estimation
Ali Akbari, Muhammad Awais, Manijeh Bashar, Josef Kittler
Good generalization performance across a wide variety of domains caused by many external and internal factors is the fundamental goal of any machine learning algorithm. This paper theoretically proves that the choice of loss function matters for improving the generalization performance of deep learning-based systems. By deriving the generalization error bound for deep neural models trained by stochastic gradient descent, we pinpoint the characteristics of the loss function that is linked to the generalization error and can therefore be used for guiding the loss function selection process. In summary, our main statement in this paper is: choose a stable loss function, generalize better. Focusing on human age estimation from the face which is a challenging topic in computer vision, we then propose a novel loss function for this learning problem. We theoretically prove that the proposed loss function achieves stronger stability, and consequently a tighter generalization error bound, compared to the other common loss functions for this problem. We have supported our findings theoretically, and demonstrated the merits of the guidance process experimentally, achieving significant improvements.
****
On Learnability via Gradient Method for Two-Layer ReLU Neural Networks in Teache

r-Student Setting
Shunta Akiyama, Taiji Suzuki
Deep learning empirically achieves high performance in many applications, but its training dynamics has not been fully understood theoretically. In this paper, we explore theoretical analysis on training two-layer ReLU neural networks in a teacher-student regression model, in which a student network learns an unknown teacher network through its outputs. We show that with a specific regularization and sufficient over-parameterization, the student network can identify the parameters of the teacher network with high probability via gradient descent with a norm dependent stepsize even though the objective function is highly non-convex. The key theoretical tool is the measure representation of the neural networks and a novel application of a dual certificate argument for sparse estimation on a measure space. We analyze the global minima and global convergence property in the measure space.
****
Slot Machines: Discovering Winning Combinations of Random Weights in Neural Networks
Maxwell M Aladago, Lorenzo Torresani
In contrast to traditional weight optimization in a continuous space, we demonstrate the existence of effective random networks whose weights are never updated. By selecting a weight among a fixed set of random values for each individual connection, our method uncovers combinations of random weights that match the performance of traditionally-trained networks of the same capacity. We refer to our networks as "slot machines" where each reel (connection) contains a fixed set of symbols (random values). Our backpropagation algorithm "spins" the reels to seek "winning" combinations, i.e., selections of random weight values that minimize the given loss. Quite surprisingly, we find that allocating just a few random values to each connection (e.g., 8 values per connection) yields highly competitive combinations despite being dramatically more constrained compared to traditionally learned weights. Moreover, finetuning these combinations often improves performance over the trained baselines. A randomly initialized VGG-19 with 8 values per connection contains a combination that achieves 91% test accuracy on CIFAR-10. Our method also achieves an impressive performance of 98.2% on MNIST for neural networks containing only random weights.
****
A large-scale benchmark for few-shot program induction and synthesis
Ferran Alet, Javier Lopez-Contreras, James Koppel, Maxwell Nye, Armando Solar-Lezama, Tomas Lozano-Perez, Leslie Kaelbling, Joshua Tenenbaum
A landmark challenge for AI is to learn flexible, powerful representations from small numbers of examples. On an important class of tasks, hypotheses in the form of programs provide extreme generalization capabilities from surprisingly few examples. However, whereas large natural few-shot learning image benchmarks have spurred progress in meta-learning for deep networks, there is no comparably big, natural program-synthesis dataset that can play a similar role. This is because, whereas images are relatively easy to label from internet meta-data or annotated by non-experts, generating meaningful input-output examples for program induction has proven hard to scale. In this work, we propose a new way of leveraging unit tests and natural inputs for small programs as meaningful input-output examples for each sub-program of the overall program. This allows us to create a large-scale naturalistic few-shot program-induction benchmark and propose new challenges in this domain. The evaluation of multiple program induction and synthesis algorithms points to shortcomings of current methods and suggests multiple avenues for future work.
****
Robust Pure Exploration in Linear Bandits with Limited Budget
Ayya Alieva, Ashok Cutkosky, Abhimanyu Das
We consider the pure exploration problem in the fixed-budget linear bandit setting. We provide a new algorithm that identifies the best arm with high probability while being robust to unknown levels of observation noise as well as to moderate levels of misspecification in the linear model. Our technique combines prior

approaches to pure exploration in the multi-armed bandit problem with optimal ex
perimental design algorithms to obtain both problem dependent and problem indepe
ndent bounds. Our success probability is never worse than that of an algorithm t
hat ignores the linear structure, but seamlessly takes advantage of such structu
re when possible. Furthermore, we only need the number of samples to scale with
the dimension of the problem rather than the number of arms. We complement our t
heoretical results with empirical validation.
****
Communication-Efficient Distributed Optimization with Quantized Preconditioners
Foivos Alimisis, Peter Davies, Dan Alistarh
We investigate fast and communication-efficient algorithms for the classic probl
em of minimizing a sum of strongly convex and smooth functions that are distribu
ted among n different nodes, which can communicate using a limited number of b
its. Most previous communication-efficient approaches for this problem are limit
ed to first-order optimization, and therefore have \emph{linear} dependence on t
he condition number in their communication complexity. We show that this depende
nce is not inherent: communication-efficient methods can in fact have sublinear
dependence on the condition number. For this, we design and analyze the first co
mmunication-efficient distributed variants of preconditioned gradient descent fo
r Generalized Linear Models, and for Newton’s method. Our results rely on a new
technique for quantizing both the preconditioner and the descent direction at ea
ch step of the algorithms, while controlling their convergence rate. We also val
idate our findings experimentally, showing faster convergence and reduced commun
ication relative to previous methods.
****
Non-Exponentially Weighted Aggregation: Regret Bounds for Unbounded Loss Functio
ns
Pierre Alquier
We tackle the problem of online optimization with a general, possibly unbounded,
 loss function. It is well known that when the loss is bounded, the exponentiall
y weighted aggregation strategy (EWA) leads to a regret in $\sqrt{T}$ after T
steps. In this paper, we study a generalized aggregation strategy, where the wei
ghts no longer depend exponentially on the losses. Our strategy is based on Foll
ow The Regularized Leader (FTRL): we minimize the expected losses plus a regular
izer, that is here a ϕ-divergence. When the regularizer is the Kullback-Lei
bler divergence, we obtain EWA as a special case. Using alternative divergences
enables unbounded losses, at the cost of a worst regret bound in some cases.
****
Dataset Dynamics via Gradient Flows in Probability Space
David Alvarez-Melis, Nicolò Fusi
Various machine learning tasks, from generative modeling to domain adaptation, r
evolve around the concept of dataset transformation and manipulation. While vari
ous methods exist for transforming unlabeled datasets, principled methods to do
so for labeled (e.g., classification) datasets are missing. In this work, we pro
pose a novel framework for dataset transformation, which we cast as optimization
 over data-generating joint probability distributions. We approach this class of
 problems through Wasserstein gradient flows in probability space, and derive pr
actical and efficient particle-based methods for a flexible but well-behaved cla
ss of objective functions. Through various experiments, we show that this framew
ork can be used to impose constraints on classification datasets, adapt them for
 transfer learning, or to re-purpose fixed or black-box models to classify {–}wi
th high accuracy{–} previously unseen datasets.
****
Submodular Maximization subject to a Knapsack Constraint: Combinatorial Algorith
ms with Near-optimal Adaptive Complexity
Georgios Amanatidis, Federico Fusco, Philip Lazos, Stefano Leonardi, Alberto Mar
chetti-Spaccamela, Rebecca Reiffenhäuser
The growing need to deal with massive instances motivates the design of algorith
ms balancing the quality of the solution with applicability. For the latter, an
important measure is the \emph{adaptive complexity}, capturing the number of seq

uential rounds of parallel computation needed. In this work we obtain the first \emph{constant factor} approximation algorithm for non-monotone submodular maximization subject to a knapsack constraint with \emph{near-optimal} $O(\log n)$ adaptive complexity. Low adaptivity by itself, however, is not enough: one needs to account for the total number of function evaluations (or value queries) as well. Our algorithm asks $\tilde{O}(n^2)$ value queries, but can be modified to run with only $\tilde{O}(n)$ instead, while retaining a low adaptive complexity of $O(\log^2n)$. Besides the above improvement in adaptivity, this is also the first \emph{combinatorial} approach with sublinear adaptive complexity for the problem and yields algorithms comparable to the state-of-the-art even for the special cases of cardinality constraints or monotone objectives. Finally, we showcase our algorithms' applicability on real-world datasets.
****
Safe Reinforcement Learning with Linear Function Approximation
Sanae Amani, Christos Thrampoulidis, Lin Yang
Safety in reinforcement learning has become increasingly important in recent years. Yet, existing solutions either fail to strictly avoid choosing unsafe actions, which may lead to catastrophic results in safety-critical systems, or fail to provide regret guarantees for settings where safety constraints need to be learned. In this paper, we address both problems by first modeling safety as an unknown linear cost function of states and actions, which must always fall below a certain threshold. We then present algorithms, termed SLUCB-QVI and RSLUCB-QVI, for episodic Markov decision processes (MDPs) with linear function approximation. We show that SLUCB-QVI and RSLUCB-QVI, while with \emph{no safety violation}, achieve a $\tilde{\mathcal{O}}\left(\kappa\sqrt{d^3H^3T}\right)$ regret, nearly matching that of state-of-the-art unsafe algorithms, where H is the duration of each episode, d is the dimension of the feature mapping, κ is a constant characterizing the safety constraints, and T is the total number of action plays. We further present numerical simulations that corroborate our theoretical findings.
****
Automatic variational inference with cascading flows
Luca Ambrogioni, Gianluigi Silvestri, Marcel van Gerven
The automation of probabilistic reasoning is one of the primary aims of machine learning. Recently, the confluence of variational inference and deep learning has led to powerful and flexible automatic inference methods that can be trained by stochastic gradient descent. In particular, normalizing flows are highly parameterized deep models that can fit arbitrarily complex posterior densities. However, normalizing flows struggle in highly structured probabilistic programs as they need to relearn the forward-pass of the program. Automatic structured variational inference (ASVI) remedies this problem by constructing variational programs that embed the forward-pass. Here, we combine the flexibility of normalizing flows and the prior-embedding property of ASVI in a new family of variational programs, which we named cascading flows. A cascading flows program interposes a newly designed highway flow architecture in between the conditional distributions of the prior program such as to steer it toward the observed data. These programs can be constructed automatically from an input probabilistic program and can also be amortized automatically. We evaluate the performance of the new variational programs in a series of structured inference problems. We find that cascading flows have much higher performance than both normalizing flows and ASVI in a large set of structured inference problems.
****
Sparse Bayesian Learning via Stepwise Regression
Sebastian E. Ament, Carla P. Gomes
Sparse Bayesian Learning (SBL) is a powerful framework for attaining sparsity in probabilistic models. Herein, we propose a coordinate ascent algorithm for SBL termed Relevance Matching Pursuit (RMP) and show that, as its noise variance parameter goes to zero, RMP exhibits a surprising connection to Stepwise Regression. Further, we derive novel guarantees for Stepwise Regression algorithms, which also shed light on RMP. Our guarantees for Forward Regression improve on determi

nistic and probabilistic results for Orthogonal Matching Pursuit with noise. Our analysis of Backward Regression culminates in a bound on the residual of the optimal solution to the subset selection problem that, if satisfied, guarantees the optimality of the result. To our knowledge, this bound is the first that can be computed in polynomial time and depends chiefly on the smallest singular value of the matrix. We report numerical experiments using a variety of feature selection algorithms. Notably, RMP and its limiting variant are both efficient and maintain strong performance with correlated features.
****
Locally Persistent Exploration in Continuous Control Tasks with Sparse Rewards
Susan Amin, Maziar Gomrokchi, Hossein Aboutalebi, Harsh Satija, Doina Precup
A major challenge in reinforcement learning is the design of exploration strategies, especially for environments with sparse reward structures and continuous state and action spaces. Intuitively, if the reinforcement signal is very scarce, the agent should rely on some form of short-term memory in order to cover its environment efficiently. We propose a new exploration method, based on two intuitions: (1) the choice of the next exploratory action should depend not only on the (Markovian) state of the environment, but also on the agent's trajectory so far, and (2) the agent should utilize a measure of spread in the state space to avoid getting stuck in a small region. Our method leverages concepts often used in statistical physics to provide explanations for the behavior of simplified (polymer) chains in order to generate persistent (locally self-avoiding) trajectories in state space. We discuss the theoretical properties of locally self-avoiding walks and their ability to provide a kind of short-term memory through a decaying temporal correlation within the trajectory. We provide empirical evaluations of our approach in a simulated 2D navigation task, as well as higher-dimensional MuJoCo continuous control locomotion tasks with sparse rewards.
****
Preferential Temporal Difference Learning
Nishanth Anand, Doina Precup
Temporal-Difference (TD) learning is a general and very useful tool for estimating the value function of a given policy, which in turn is required to find good policies. Generally speaking, TD learning updates states whenever they are visited. When the agent lands in a state, its value can be used to compute the TD-error, which is then propagated to other states. However, it may be interesting, when computing updates, to take into account other information than whether a state is visited or not. For example, some states might be more important than others (such as states which are frequently seen in a successful trajectory). Or, some states might have unreliable value estimates (for example, due to partial observability or lack of data), making their values less desirable as targets. We propose an approach to re-weighting states used in TD updates, both when they are the input and when they provide the target for the update. We prove that our approach converges with linear function approximation and illustrate its desirable empirical behaviour compared to other TD-style methods.
****
Unitary Branching Programs: Learnability and Lower Bounds
Fidel Ernesto Diaz Andino, Maria Kokkou, Mateus De Oliveira Oliveira, Farhad Vadiee
Bounded width branching programs are a formalism that can be used to capture the notion of non-uniform constant-space computation. In this work, we study a generalized version of bounded width branching programs where instructions are defined by unitary matrices of bounded dimension. We introduce a new learning framework for these branching programs that leverages on a combination of local search techniques with gradient descent over Riemannian manifolds. We also show that gapped, read-once branching programs of bounded dimension can be learned with a polynomial number of queries in the presence of a teacher. Finally, we provide explicit near-quadratic size lower-bounds for bounded-dimension unitary branching programs, and exponential size lower-bounds for bounded-dimension read-once gapped unitary branching programs. The first lower bound is proven using a combination of Neciporuk's lower bound technique with classic results from algebraic geome

try. The second lower bound is proven within the framework of communication comp
lexity theory.
****
The Logical Options Framework
Brandon Araki, Xiao Li, Kiran Vodrahalli, Jonathan Decastro, Micah Fry, Daniela
Rus
Learning composable policies for environments with complex rules and tasks is a
challenging problem. We introduce a hierarchical reinforcement learning framewor
k called the Logical Options Framework (LOF) that learns policies that are satis
fying, optimal, and composable. LOF efficiently learns policies that satisfy tas
ks by representing the task as an automaton and integrating it into learning and
 planning. We provide and prove conditions under which LOF will learn satisfying
, optimal policies. And lastly, we show how LOF’s learned policies can be compos
ed to satisfy unseen tasks with only 10-50 retraining steps on our benchmarks. W
e evaluate LOF on four tasks in discrete and continuous domains, including a 3D
pick-and-place environment.
****
Annealed Flow Transport Monte Carlo
Michael Arbel, Alex Matthews, Arnaud Doucet
Annealed Importance Sampling (AIS) and its Sequential Monte Carlo (SMC) extensio
ns are state-of-the-art methods for estimating normalizing constants of probabil
ity distributions. We propose here a novel Monte Carlo algorithm, Annealed Flow
Transport (AFT), that builds upon AIS and SMC and combines them with normalizing
 flows (NFs) for improved performance. This method transports a set of particles
 using not only importance sampling (IS), Markov chain Monte Carlo (MCMC) and re
sampling steps - as in SMC, but also relies on NFs which are learned sequentiall
y to push particles towards the successive annealed targets. We provide limit th
eorems for the resulting Monte Carlo estimates of the normalizing constant and e
xpectations with respect to the target distribution. Additionally, we show that
a continuous-time scaling limit of the population version of AFT is given by a F
eynman–Kac measure which simplifies to the law of a controlled diffusion for exp
ressive NFs. We demonstrate experimentally the benefits and limitations of our m
ethodology on a variety of applications.
****
Permutation Weighting
David Arbour, Drew Dimmery, Arjun Sondhi
A commonly applied approach for estimating causal effects from observational dat
a is to apply weights which render treatments independent of observed pre-treatm
ent covariates. Recently emphasis has been placed on deriving balancing weights
which explicitly target this independence condition. In this work we introduce p
ermutation weighting, a method for estimating balancing weights using a standard
 binary classifier (regardless of cardinality of treatment). A large class of pr
obabilistic classifiers may be used in this method; the choice of loss for the c
lassifier implies the particular definition of balance. We bound bias and varian
ce in terms of the excess risk of the classifier, show that these disappear asym
ptotically, and demonstrate that our classification problem directly minimizes i
mbalance. Additionally, hyper-parameter tuning and model selection can be perfor
med with standard cross-validation methods. Empirical evaluations indicate that
permutation weighting provides favorable performance in comparison to existing m
ethods.
****
Analyzing the tree-layer structure of Deep Forests
Ludovic Arnould, Claire Boyer, Erwan Scornet
Random forests on the one hand, and neural networks on the other hand, have met
great success in the machine learning community for their predictive performance
. Combinations of both have been proposed in the literature, notably leading to
the so-called deep forests (DF) (Zhou & Feng,2019). In this paper, our aim is no
t to benchmark DF performances but to investigate instead their underlying mecha
nisms. Additionally, DF architecture can be generally simplified into more simpl
e and computationally efficient shallow forest networks. Despite some instabilit

y, the latter may outperform standard predictive tree-based methods. We exhibit a theoretical framework in which a shallow tree network is shown to enhance the performance of classical decision trees. In such a setting, we provide tight theoretical lower and upper bounds on its excess risk. These theoretical results show the interest of tree-network architectures for well-structured data provided that the first layer, acting as a data encoder, is rich enough.
****

Dropout: Explicit Forms and Capacity Control

Raman Arora, Peter Bartlett, Poorya Mianjy, Nathan Srebro

We investigate the capacity control provided by dropout in various machine learning problems. First, we study dropout for matrix completion, where it induces a distribution-dependent regularizer that equals the weighted trace-norm of the product of the factors. In deep learning, we show that the distribution-dependent regularizer due to dropout directly controls the Rademacher complexity of the underlying class of deep neural networks. These developments enable us to give concrete generalization error bounds for the dropout algorithm in both matrix completion as well as training deep neural networks.
****

Tighter Bounds on the Log Marginal Likelihood of Gaussian Process Regression Using Conjugate Gradients

Artem Artemev, David R. Burt, Mark van der Wilk

We propose a lower bound on the log marginal likelihood of Gaussian process regression models that can be computed without matrix factorisation of the full kernel matrix. We show that approximate maximum likelihood learning of model parameters by maximising our lower bound retains many benefits of the sparse variational approach while reducing the bias introduced into hyperparameter learning. The basis of our bound is a more careful analysis of the log-determinant term appearing in the log marginal likelihood, as well as using the method of conjugate gradients to derive tight lower bounds on the term involving a quadratic form. Our approach is a step forward in unifying methods relying on lower bound maximisation (e.g. variational methods) and iterative approaches based on conjugate gradients for training Gaussian processes. In experiments, we show improved predictive performance with our model for a comparable amount of training time compared to other conjugate gradient based approaches.
****

Deciding What to Learn: A Rate-Distortion Approach

Dilip Arumugam, Benjamin Van Roy

Agents that learn to select optimal actions represent a prominent focus of the sequential decision-making literature. In the face of a complex environment or constraints on time and resources, however, aiming to synthesize such an optimal policy can become infeasible. These scenarios give rise to an important trade-off between the information an agent must acquire to learn and the sub-optimality of the resulting policy. While an agent designer has a preference for how this trade-off is resolved, existing approaches further require that the designer translate these preferences into a fixed learning target for the agent. In this work, leveraging rate-distortion theory, we automate this process such that the designer need only express their preferences via a single hyperparameter and the agent is endowed with the ability to compute its own learning targets that best achieve the desired trade-off. We establish a general bound on expected discounted regret for an agent that decides what to learn in this manner along with computational experiments that illustrate the expressiveness of designer preferences and even show improvements over Thompson sampling in identifying an optimal policy.
****

Private Adaptive Gradient Methods for Convex Optimization

Hilal Asi, John Duchi, Alireza Fallah, Omid Javidbakht, Kunal Talwar

We study adaptive methods for differentially private convex optimization, proposing and analyzing differentially private variants of a Stochastic Gradient Descent (SGD) algorithm with adaptive stepsizes, as well as the AdaGrad algorithm. We provide upper bounds on the regret of both algorithms and show that the bounds are (worst-case) optimal. As a consequence of our development, we show that our

private versions of AdaGrad outperform adaptive SGD, which in turn outperforms t
raditional SGD in scenarios with non-isotropic gradients where (non-private) Ada
grad provably outperforms SGD. The major challenge is that the isotropic noise t
ypically added for privacy dominates the signal in gradient geometry for high-di
mensional problems; approaches to this that effectively optimize over lower-dime
nsional subspaces simply ignore the actual problems that varying gradient geomet
ries introduce. In contrast, we study non-isotropic clipping and noise addition,
 developing a principled theoretical approach; the consequent procedures also en
joy significantly stronger empirical performance than prior approaches.
****
Private Stochastic Convex Optimization: Optimal Rates in L1 Geometry
Hilal Asi, Vitaly Feldman, Tomer Koren, Kunal Talwar
Stochastic convex optimization over an ℓ_1-bounded domain is ubiquitous in
machine learning applications such as LASSO but remains poorly understood when l
earning with differential privacy. We show that, up to logarithmic factors the o
ptimal excess population loss of any (ϵ,δ)-differentially private
optimizer is $\sqrt{\log(d)/n} + \sqrt{d}/\epsilon n.$ The upper bound is based
on a new algorithm that combines the iterative localization approach of Feldman
et al. (2020) with a new analysis of private regularized mirror descent. It appl
ies to ℓ_p bounded domains for $p\in [1,2]$ and queries at most $n^{3/2}$ g
radients improving over the best previously known algorithm for the ℓ_2 cas
e which needs n^2 gradients. Further, we show that when the loss functions sat
isfy additional smoothness assumptions, the excess loss is upper bounded (up to
logarithmic factors) by $\sqrt{\log(d)/n} + (\log(d)/\epsilon n)^{2/3}.$ This bo
und is achieved by a new variance-reduced version of the Frank-Wolfe algorithm t
hat requires just a single pass over the data. We also show that the lower bound
 in this case is the minimum of the two rates mentioned above.
****
Combinatorial Blocking Bandits with Stochastic Delays
Alexia Atsidakou, Orestis Papadigenopoulos, Soumya Basu, Constantine Caramanis,
Sanjay Shakkottai
Recent work has considered natural variations of the {\em multi-armed bandit} pr
oblem, where the reward distribution of each arm is a special function of the ti
me passed since its last pulling. In this direction, a simple (yet widely applic
able) model is that of {\em blocking bandits}, where an arm becomes unavailable
for a deterministic number of rounds after each play. In this work, we extend th
e above model in two directions: (i) We consider the general combinatorial setti
ng where more than one arms can be played at each round, subject to feasibility
constraints. (ii) We allow the blocking time of each arm to be stochastic. We fi
rst study the computational/unconditional hardness of the above setting and iden
tify the necessary conditions for the problem to become tractable (even in an ap
proximate sense). Based on these conditions, we provide a tight analysis of the
approximation guarantee of a natural greedy heuristic that always plays the maxi
mum expected reward feasible subset among the available (non-blocked) arms. When
 the arms’ expected rewards are unknown, we adapt the above heuristic into a ban
dit algorithm, based on UCB, for which we provide sublinear (approximate) regret
 guarantees, matching the theoretical lower bounds in the limiting case of absen
ce of delays.
****
Dichotomous Optimistic Search to Quantify Human Perception
Julien Audiffren
In this paper we address a variant of the continuous multi-armed bandits problem
, called the threshold estimation problem, which is at the heart of many psychom
etric experiments. Here, the objective is to estimate the sensitivity threshold
for an unknown psychometric function Psi, which is assumed to be non decreasing
and continuous. Our algorithm, Dichotomous Optimistic Search (DOS), efficiently
solves this task by taking inspiration from hierarchical multi-armed bandits and
 Black-box optimization. Compared to previous approaches, DOS is model free and
only makes minimal assumption on Psi smoothness, while having strong theoretical
 guarantees that compares favorably to recent methods from both Psychophysics an

d Global Optimization. We also empirically evaluate DOS and show that it signifi cantly outperforms these methods, both in experiments that mimics the conduct of a psychometric experiment, and in tests with large pulls budgets that illustrat e the faster convergence rate.
****

Federated Learning under Arbitrary Communication Patterns
Dmitrii Avdiukhin, Shiva Kasiviswanathan
Federated Learning is a distributed learning setting where the goal is to train a centralized model with training data distributed over a large number of hetero geneous clients, each with unreliable and relatively slow network connections. A common optimization approach used in federated learning is based on the idea of local SGD: each client runs some number of SGD steps locally and then the updat ed local models are averaged to form the updated global model on the coordinatin g server. In this paper, we investigate the performance of an asynchronous versi on of local SGD wherein the clients can communicate with the server at arbitrary time intervals. Our main result shows that for smooth strongly convex and smoot h nonconvex functions we achieve convergence rates that match the synchronous ve rsion that requires all clients to communicate simultaneously.
****

Asynchronous Distributed Learning : Adapting to Gradient Delays without Prior Kn owledge
Rotem Zamir Aviv, Ido Hakimi, Assaf Schuster, Kfir Yehuda Levy
We consider stochastic convex optimization problems, where several machines act asynchronously in parallel while sharing a common memory. We propose a robust tr aining method for the constrained setting and derive non asymptotic convergence guarantees that do not depend on prior knowledge of update delays, objective smo othness, and gradient variance. Conversely, existing methods for this setting cr ucially rely on this prior knowledge, which render them unsuitable for essential ly all shared-resources computational environments, such as clouds and data cent ers. Concretely, existing approaches are unable to accommodate changes in the de lays which result from dynamic allocation of the machines, while our method impl icitly adapts to such changes.
****

Decomposable Submodular Function Minimization via Maximum Flow
Kyriakos Axiotis, Adam Karczmarz, Anish Mukherjee, Piotr Sankowski, Adrian Vladu
This paper bridges discrete and continuous optimization approaches for decomposa ble submodular function minimization, in both the standard and parametric settin gs. We provide improved running times for this problem by reducing it to a numbe r of calls to a maximum flow oracle. When each function in the decomposition act s on O(1) elements of the ground set V and is polynomially bounded, our running time is up to polylogarithmic factors equal to that of solving maximum flow in a sparse graph with O(|V|) vertices and polynomial integral capacities. We achiev e this by providing a simple iterative method which can optimize to high precisi on any convex function defined on the submodular base polytope, provided we can efficiently minimize it on the base polytope corresponding to the cut function o f a certain graph that we construct. We solve this minimization problem by lifti ng the solutions of a parametric cut problem, which we obtain via a new efficien t combinatorial reduction to maximum flow. This reduction is of independent inte rest and implies some previously unknown bounds for the parametric minimum s,t-c ut problem in multiple settings.
****

Differentially Private Query Release Through Adaptive Projection
Sergul Aydore, William Brown, Michael Kearns, Krishnaram Kenthapadi, Luca Melis, Aaron Roth, Ankit A. Siva
We propose, implement, and evaluate a new algo-rithm for releasing answers to ve ry large numbersof statistical queries likek-way marginals, sub-ject to differen tial privacy. Our algorithm makesadaptive use of a continuous relaxation of theP ro-jection Mechanism, which answers queries on theprivate dataset using simple p erturbation, and thenattempts to find the synthetic dataset that mostclosely mat ches the noisy answers. We use a con-tinuous relaxation of the synthetic dataset

domainwhich makes the projection loss differentiable,and allows us to use efficient ML optimizationtechniques and tooling. Rather than answering allqueries up front, we make judicious use of ourprivacy budget by iteratively finding queries forwhich our (relaxed) synthetic data has high error,and then repeating the projection. Randomizedrounding allows us to obtain synthetic data in theoriginal schema. We perform experimental evalu-ations across a range of parameters and datasets,and find that our method outperforms existingalgorithms on large query classes.
****
On the Implicit Bias of Initialization Shape: Beyond Infinitesimal Mirror Descent
Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E Woodworth, Nathan Srebro, Amir Globerson, Daniel Soudry
Recent work has highlighted the role of initialization scale in determining the structure of the solutions that gradient methods converge to. In particular, it was shown that large initialization leads to the neural tangent kernel regime solution, whereas small initialization leads to so called “rich regimes”. However, the initialization structure is richer than the overall scale alone and involves relative magnitudes of different weights and layers in the network. Here we show that these relative scales, which we refer to as initialization shape, play an important role in determining the learned model. We develop a novel technique for deriving the inductive bias of gradient-flow and use it to obtain closed-form implicit regularizers for multiple cases of interest.
****
On-Off Center-Surround Receptive Fields for Accurate and Robust Image Classification
Zahra Babaiee, Ramin Hasani, Mathias Lechner, Daniela Rus, Radu Grosu
Robustness to variations in lighting conditions is a key objective for any deep vision system. To this end, our paper extends the receptive field of convolutional neural networks with two residual components, ubiquitous in the visual processing system of vertebrates: On-center and off-center pathways, with an excitatory center and inhibitory surround; OOCS for short. The On-center pathway is excited by the presence of a light stimulus in its center, but not in its surround, whereas the Off-center pathway is excited by the absence of a light stimulus in its center, but not in its surround. We design OOCS pathways via a difference of Gaussians, with their variance computed analytically from the size of the receptive fields. OOCS pathways complement each other in their response to light stimuli, ensuring this way a strong edge-detection capability, and as a result an accurate and robust inference under challenging lighting conditions. We provide extensive empirical evidence showing that networks supplied with OOCS pathways gain accuracy and illumination-robustness from the novel edge representation, compared to other baselines.
****
Uniform Convergence, Adversarial Spheres and a Simple Remedy
Gregor Bachmann, Seyed-Mohsen Moosavi-Dezfooli, Thomas Hofmann
Previous work has cast doubt on the general framework of uniform convergence and its ability to explain generalization in neural networks. By considering a specific dataset, it was observed that a neural network completely misclassifies a projection of the training data (adversarial set), rendering any existing generalization bound based on uniform convergence vacuous. We provide an extensive theoretical investigation of the previously studied data setting through the lens of infinitely-wide models. We prove that the Neural Tangent Kernel (NTK) also suffers from the same phenomenon and we uncover its origin. We highlight the important role of the output bias and show theoretically as well as empirically how a sensible choice completely mitigates the problem. We identify sharp phase transitions in the accuracy on the adversarial set and study its dependency on the training sample size. As a result, we are able to characterize critical sample sizes beyond which the effect disappears. Moreover, we study decompositions of a neural network into a clean and noisy part by considering its canonical decomposition into its different eigenfunctions and show empirically that for too small bias

the adversarial phenomenon still persists.
****
Faster Kernel Matrix Algebra via Density Estimation
Arturs Backurs, Piotr Indyk, Cameron Musco, Tal Wagner
We study fast algorithms for computing basic properties of an n x n positive semidefinite kernel matrix K corresponding to n points x_1,...,x_n in R^d. In particular, we consider the estimating the sum of kernel matrix entries, along with its top eigenvalue and eigenvector. These are some of the most basic problems defined over kernel matrices. We show that the sum of matrix entries can be estimated up to a multiplicative factor of 1+\epsilon in time sublinear in n and linear in d for many popular kernel functions, including the Gaussian, exponential, and rational quadratic kernels. For these kernels, we also show that the top eigenvalue (and a witnessing approximate eigenvector) can be approximated to a multiplicative factor of 1+\epsilon in time sub-quadratic in n and linear in d. Our algorithms represent significant advances in the best known runtimes for these problems. They leverage the positive definiteness of the kernel matrix, along with a recent line of work on efficient kernel density estimation.
****
Robust Reinforcement Learning using Least Squares Policy Iteration with Provable Performance Guarantees
Kishan Panaganti Badrinath, Dileep Kalathil
This paper addresses the problem of model-free reinforcement learning for Robust Markov Decision Process (RMDP) with large state spaces. The goal of the RMDPs framework is to find a policy that is robust against the parameter uncertainties due to the mismatch between the simulator model and real-world settings. We first propose the Robust Least Squares Policy Evaluation algorithm, which is a multi-step online model-free learning algorithm for policy evaluation. We prove the convergence of this algorithm using stochastic approximation techniques. We then propose Robust Least Squares Policy Iteration (RLSPI) algorithm for learning the optimal robust policy. We also give a general weighted Euclidean norm bound on the error (closeness to optimality) of the resulting policy. Finally, we demonstrate the performance of our RLSPI algorithm on some benchmark problems from OpenAI Gym.
****
Skill Discovery for Exploration and Planning using Deep Skill Graphs
Akhil Bagaria, Jason K Senthil, George Konidaris
We introduce a new skill-discovery algorithm that builds a discrete graph representation of large continuous MDPs, where nodes correspond to skill subgoals and the edges to skill policies. The agent constructs this graph during an unsupervised training phase where it interleaves discovering skills and planning using them to gain coverage over ever-increasing portions of the state-space. Given a novel goal at test time, the agent plans with the acquired skill graph to reach a nearby state, then switches to learning to reach the goal. We show that the resulting algorithm, Deep Skill Graphs, outperforms both flat and existing hierarchical reinforcement learning methods on four difficult continuous control tasks.
****
Locally Adaptive Label Smoothing Improves Predictive Churn
Dara Bahri, Heinrich Jiang
Training modern neural networks is an inherently noisy process that can lead to high \emph{prediction churn}– disagreements between re-trainings of the same model due to factors such as randomization in the parameter initialization and mini-batches– even when the trained models all attain similar accuracies. Such prediction churn can be very undesirable in practice. In this paper, we present several baselines for reducing churn and show that training on soft labels obtained by adaptively smoothing each example’s label based on the example’s neighboring labels often outperforms the baselines on churn while improving accuracy on a variety of benchmark classification tasks and model architectures.
****
How Important is the Train-Validation Split in Meta-Learning?
Yu Bai, Minshuo Chen, Pan Zhou, Tuo Zhao, Jason Lee, Sham Kakade, Huan Wang, Cai

ming Xiong
Meta-learning aims to perform fast adaptation on a new task through learning a “prior” from multiple existing tasks. A common practice in meta-learning is to perform a train-validation split (\emph{train-val method}) where the prior adapts to the task on one split of the data, and the resulting predictor is evaluated on another split. Despite its prevalence, the importance of the train-validation split is not well understood either in theory or in practice, particularly in comparison to the more direct \emph{train-train method}, which uses all the per-task data for both training and evaluation. We provide a detailed theoretical study on whether and when the train-validation split is helpful in the linear centroid meta-learning problem. In the agnostic case, we show that the expected loss of the train-val method is minimized at the optimal prior for meta testing, and this is not the case for the train-train method in general without structural assumptions on the data. In contrast, in the realizable case where the data are generated from linear models, we show that both the train-val and train-train losses are minimized at the optimal prior in expectation. Further, perhaps surprisingly, our main result shows that the train-train method achieves a \emph{strictly better} excess loss in this realizable case, even when the regularization parameter and split ratio are optimally tuned for both methods. Our results highlight that sample splitting may not always be preferable, especially when the data is realizable by the model. We validate our theories by experimentally showing that the train-train method can indeed outperform the train-val method, on both simulations and real meta-learning tasks.
****
Stabilizing Equilibrium Models by Jacobian Regularization
Shaojie Bai, Vladlen Koltun, Zico Kolter
Deep equilibrium networks (DEQs) are a new class of models that eschews traditional depth in favor of finding the fixed point of a single non-linear layer. These models have been shown to achieve performance competitive with the state-of-the-art deep networks while using significantly less memory. Yet they are also slower, brittle to architectural choices, and introduce potential instability to the model. In this paper, we propose a regularization scheme for DEQ models that explicitly regularizes the Jacobian of the fixed-point update equations to stabilize the learning of equilibrium models. We show that this regularization adds only minimal computational cost, significantly stabilizes the fixed-point convergence in both forward and backward passes, and scales well to high-dimensional, realistic domains (e.g., WikiText-103 language modeling and ImageNet classification). Using this method, we demonstrate, for the first time, an implicit-depth model that runs with approximately the same speed and level of performance as popular conventional deep networks such as ResNet-101, while still maintaining the constant memory footprint and architectural simplicity of DEQs. Code is available https://github.com/locuslab/deq.
****
Don’t Just Blame Over-parametrization for Over-confidence: Theoretical Analysis of Calibration in Binary Classification
Yu Bai, Song Mei, Huan Wang, Caiming Xiong
Modern machine learning models with high accuracy are often miscalibrated—the predicted top probability does not reflect the actual accuracy, and tends to be \emph{over-confident}. It is commonly believed that such over-confidence is mainly due to \emph{over-parametrization}, in particular when the model is large enough to memorize the training data and maximize the confidence. In this paper, we show theoretically that over-parametrization is not the only reason for over-confidence. We prove that \emph{logistic regression is inherently over-confident}, in the realizable, under-parametrized setting where the data is generated from the logistic model, and the sample size is much larger than the number of parameters. Further, this over-confidence happens for general well-specified binary classification problems as long as the activation is symmetric and concave on the positive part. Perhaps surprisingly, we also show that over-confidence is not always the case—there exists another activation function (and a suitable loss function) under which the learned classifier is \emph{under-confident} at some probabi

lity values. Overall, our theory provides a precise characterization of calibration in realizable binary classification, which we verify on simulations and real data experiments.
****
Principled Exploration via Optimistic Bootstrapping and Backward Induction
Chenjia Bai, Lingxiao Wang, Lei Han, Jianye Hao, Animesh Garg, Peng Liu, Zhaoran Wang
One principled approach for provably efficient exploration is incorporating the upper confidence bound (UCB) into the value function as a bonus. However, UCB is specified to deal with linear and tabular settings and is incompatible with Deep Reinforcement Learning (DRL). In this paper, we propose a principled exploration method for DRL through Optimistic Bootstrapping and Backward Induction (OB2I). OB2I constructs a general-purpose UCB-bonus through non-parametric bootstrap in DRL. The UCB-bonus estimates the epistemic uncertainty of state-action pairs for optimistic exploration. We build theoretical connections between the proposed UCB-bonus and the LSVI-UCB in linear setting. We propagate future uncertainty in a time-consistent manner through episodic backward update, which exploits the theoretical advantage and empirically improves the sample-efficiency. Our experiments in MNIST maze and Atari suit suggest that OB2I outperforms several state-of-the-art exploration approaches.
****
GLSearch: Maximum Common Subgraph Detection via Learning to Search
Yunsheng Bai, Derek Xu, Yizhou Sun, Wei Wang
Detecting the Maximum Common Subgraph (MCS) between two input graphs is fundamental for applications in drug synthesis, malware detection, cloud computing, etc. However, MCS computation is NP-hard, and state-of-the-art MCS solvers rely on heuristic search algorithms which in practice cannot find good solution for large graph pairs given a limited computation budget. We propose GLSearch, a Graph Neural Network (GNN) based learning to search model. Our model is built upon the branch and bound algorithm, which selects one pair of nodes from the two input graphs to expand at a time. We propose a novel GNN-based Deep Q-Network (DQN) to select the node pair, making the search process much faster. Experiments on synthetic and real-world graph pairs demonstrate that our model learns a search strategy that is able to detect significantly larger common subgraphs than existing MCS solvers given the same computation budget. GLSearch can be potentially extended to solve many other combinatorial problems with constraints on graphs.
****
Breaking the Limits of Message Passing Graph Neural Networks
Muhammet Balcilar, Pierre Heroux, Benoit Gauzere, Pascal Vasseur, Sebastien Adam, Paul Honeine
Since the Message Passing (Graph) Neural Networks (MPNNs) have a linear complexity with respect to the number of nodes when applied to sparse graphs, they have been widely implemented and still raise a lot of interest even though their theoretical expressive power is limited to the first order Weisfeiler-Lehman test (1-WL). In this paper, we show that if the graph convolution supports are designed in spectral-domain by a non-linear custom function of eigenvalues and masked with an arbitrary large receptive field, the MPNN is theoretically more powerful than the 1-WL test and experimentally as powerful as a 3-WL existing models, while remaining spatially localized. Moreover, by designing custom filter functions, outputs can have various frequency components that allow the convolution process to learn different relationships between a given input graph signal and its associated properties. So far, the best 3-WL equivalent graph neural networks have a computational complexity in $\mathcal{O}(n^3)$ with memory usage in $\mathcal{O}(n^2)$, consider non-local update mechanism and do not provide the spectral richness of output profile. The proposed method overcomes all these aforementioned problems and reaches state-of-the-art results in many downstream tasks.
****
Instance Specific Approximations for Submodular Maximization
Eric Balkanski, Sharon Qian, Yaron Singer
The predominant measure for the performance of an algorithm is its worst-case ap

proximation guarantee. While worst-case approximations give desirable robustness
 guarantees, they can differ significantly from the performance of an algorithm
in practice. For the problem of monotone submodular maximization under a cardina
lity constraint, the greedy algorithm is known to obtain a 1-1/e approximation g
uarantee, which is optimal for a polynomial-time algorithm. However, very little
 is known about the approximation achieved by greedy and other submodular maximi
zation algorithms on real instances. We develop an algorithm that gives an insta
nce-specific approximation for any solution of an instance of monotone submodula
r maximization under a cardinality constraint. This algorithm uses a novel dual
approach to submodular maximization. In particular, it relies on the constructio
n of a lower bound to the dual objective that can also be exactly minimized. We
use this algorithm to show that on a wide variety of real-world datasets and obj
ectives, greedy and other algorithms find solutions that approximate the optimal
 solution significantly better than the 1-1/e   0.63 worst-case approximation gu
arantee, often exceeding 0.9.
****
Augmented World Models Facilitate Zero-Shot Dynamics Generalization From a Singl
e Offline Environment
Philip J Ball, Cong Lu, Jack Parker-Holder, Stephen Roberts
Reinforcement learning from large-scale offline datasets provides us with the ab
ility to learn policies without potentially unsafe or impractical exploration. S
ignificant progress has been made in the past few years in dealing with the chal
lenge of correcting for differing behavior between the data collection and learn
ed policies. However, little attention has been paid to potentially changing dyn
amics when transferring a policy to the online setting, where performance can be
 up to 90% reduced for existing methods. In this paper we address this problem w
ith Augmented World Models (AugWM). We augment a learned dynamics model with sim
ple transformations that seek to capture potential changes in physical propertie
s of the robot, leading to more robust policies. We not only train our policy in
 this new setting, but also provide it with the sampled augmentation as a contex
t, allowing it to adapt to changes in the environment. At test time we learn the
 context in a self-supervised fashion by approximating the augmentation which co
rresponds to the new environment. We rigorously evaluate our approach on over 10
0 different changed dynamics settings, and show that this simple approach can si
gnificantly improve the zero-shot generalization of a recent state-of-the-art ba
seline, often achieving successful policies where the baseline fails.
****
Regularized Online Allocation Problems: Fairness and Beyond
Santiago Balseiro, Haihao Lu, Vahab Mirrokni
Online allocation problems with resource constraints have a rich history in comp
uter science and operations research. In this paper, we introduce the regularize
d online allocation problem, a variant that includes a non-linear regularizer ac
ting on the total resource consumption. In this problem, requests repeatedly arr
ive over time and, for each request, a decision maker needs to take an action th
at generates a reward and consumes resources. The objective is to simultaneously
 maximize total rewards and the value of the regularizer subject to the resource
 constraints. Our primary motivation is the online allocation of internet advert
isements wherein firms seek to maximize additive objectives such as the revenue
or efficiency of the allocation. By introducing a regularizer, firms can account
 for the fairness of the allocation or, alternatively, punish under-delivery of
advertisements—two common desiderata in internet advertising markets. We design
an algorithm when arrivals are drawn independently from a distribution that is u
nknown to the decision maker. Our algorithm is simple, fast, and attains the opt
imal order of sub-linear regret compared to the optimal allocation with the bene
fit of hindsight. Numerical experiments confirm the effectiveness of the propose
d algorithm and of the regularizers in an internet advertising application.
****
Predict then Interpolate: A Simple Algorithm to Learn Stable Classifiers
Yujia Bao, Shiyu Chang, Regina Barzilay
We propose Predict then Interpolate (PI), a simple algorithm for learning correl

ations that are stable across environments. The algorithm follows from the intui tion that when using a classifier trained on one environment to make predictions on examples from another environment, its mistakes are informative as to which correlations are unstable. In this work, we prove that by interpolating the dist ributions of the correct predictions and the wrong predictions, we can uncover a n oracle distribution where the unstable correlation vanishes. Since the oracle interpolation coefficients are not accessible, we use group distributionally rob ust optimization to minimize the worst-case risk across all such interpolations. We evaluate our method on both text classification and image classification. Em pirical results demonstrate that our algorithm is able to learn robust classifie rs (outperforms IRM by 23.85% on synthetic environments and 12.41% on natural en vironments). Our code and data are available at https://github.com/YujiaBao/ Pre dict-then-Interpolate.
****
Variational (Gradient) Estimate of the Score Function in Energy-based Latent Var iable Models
Fan Bao, Kun Xu, Chongxuan Li, Lanqing Hong, Jun Zhu, Bo Zhang
This paper presents new estimates of the score function and its gradient with re spect to the model parameters in a general energy-based latent variable model (E BLVM). The score function and its gradient can be expressed as combinations of e xpectation and covariance terms over the (generally intractable) posterior of th e latent variables. New estimates are obtained by introducing a variational post erior to approximate the true posterior in these terms. The variational posterio r is trained to minimize a certain divergence (e.g., the KL divergence) between itself and the true posterior. Theoretically, the divergence characterizes upper bounds of the bias of the estimates. In principle, our estimates can be applied to a wide range of objectives, including kernelized Stein discrepancy (KSD), sc ore matching (SM)-based methods and exact Fisher divergence with a minimal model assumption. In particular, these estimates applied to SM-based methods outperfo rm existing methods in learning EBLVMs on several image datasets.
****
Compositional Video Synthesis with Action Graphs
Amir Bar, Roei Herzig, Xiaolong Wang, Anna Rohrbach, Gal Chechik, Trevor Darrell , Amir Globerson
Videos of actions are complex signals containing rich compositional structure in space and time. Current video generation methods lack the ability to condition the generation on multiple coordinated and potentially simultaneous timed action s. To address this challenge, we propose to represent the actions in a graph str ucture called Action Graph and present the new "Action Graph To Video" synthesis task. Our generative model for this task (AG2Vid) disentangles motion and appea rance features, and by incorporating a scheduling mechanism for actions facilita tes a timely and coordinated video generation. We train and evaluate AG2Vid on C ATER and Something-Something V2 datasets, which results in videos that have bett er visual quality and semantic consistency compared to baselines. Finally, our m odel demonstrates zero-shot abilities by synthesizing novel compositions of the learned actions.
****
Approximating a Distribution Using Weight Queries
Nadav Barak, Sivan Sabato
We consider a novel challenge: approximating a distribution without the ability to randomly sample from that distribution. We study how such an approximation ca n be obtained using *weight queries*. Given some data set of examples, a weight query presents one of the examples to an oracle, which returns the probability, according to the target distribution, of observing examples similar to the prese nted example. This oracle can represent, for instance, counting queries to a dat abase of the target population, or an interface to a search engine which returns the number of results that match a given search. We propose an interactive algo rithm that iteratively selects data set examples and performs corresponding weig ht queries. The algorithm finds a reweighting of the data set that approximates the weights according to the target distribution, using a limited number of weig

ht queries. We derive an approximation bound on the total variation distance bet
ween the reweighting found by the algorithm and the best achievable reweighting.
 Our algorithm takes inspiration from the UCB approach common in multi-armed ban
dits problems, and combines it with a new discrepancy estimator and a greedy ite
rative procedure. In addition to our theoretical guarantees, we demonstrate in e
xperiments the advantages of the proposed algorithm over several baselines. A py
thon implementation of the proposed algorithm and of all the experiments can be
found at https://github.com/Nadav-Barak/AWP.
****
Graph Convolution for Semi-Supervised Classification: Improved Linear Separabili
ty and Out-of-Distribution Generalization
Aseem Baranwal, Kimon Fountoulakis, Aukosh Jagannath
Recently there has been increased interest in semi-supervised classification in
the presence of graphical information. A new class of learning models has emerge
d that relies, at its most basic level, on classifying the data after first appl
ying a graph convolution. To understand the merits of this approach, we study th
e classification of a mixture of Gaussians, where the data corresponds to the no
de attributes of a stochastic block model. We show that graph convolution extend
s the regime in which the data is linearly separable by a factor of roughly $1/\
sqrt{D}$, where D is the expected degree of a node, as compared to the mixture
 model data on its own. Furthermore, we find that the linear classifier obtained
 by minimizing the cross-entropy loss after the graph convolution generalizes to
 out-of-distribution data where the unseen data can have different intra- and in
ter-class edge probabilities from the training data.
****
Training Quantized Neural Networks to Global Optimality via Semidefinite Program
ming
Burak Bartan, Mert Pilanci
Neural networks (NNs) have been extremely successful across many tasks in machin
e learning. Quantization of NN weights has become an important topic due to its
impact on their energy efficiency, inference time and deployment on hardware. Al
though post-training quantization is well-studied, training optimal quantized NN
s involves combinatorial non-convex optimization problems which appear intractab
le. In this work, we introduce a convex optimization strategy to train quantized
 NNs with polynomial activations. Our method leverages hidden convexity in two-l
ayer neural networks from the recent literature, semidefinite lifting, and Groth
endieck’s identity. Surprisingly, we show that certain quantized NN problems can
 be solved to global optimality provably in polynomial time in all relevant para
meters via tight semidefinite relaxations. We present numerical examples to illu
strate the effectiveness of our method.
****
Beyond $log^2(T)$ regret for decentralized bandits in matching markets
Soumya Basu, Karthik Abinav Sankararaman, Abishek Sankararaman
We design decentralized algorithms for regret minimization in the two sided matc
hing market with one-sided bandit feedback that significantly improves upon the
prior works (Liu et al.\,2020a, Sankararaman et al.\,2020, Liu et al.\,2020b). F
irst, for general markets, for any $\varepsilon > 0$, we design an algorithm tha
t achieves a $O(\log^{1+\varepsilon}(T))$ regret to the agent-optimal stable mat
ching, with unknown time horizon T, improving upon the $O(\log^{2}(T))$ regret
 achieved in (Liu et al.\,2020b). Second, we provide the optimal $\Theta(\log(T)
)$ agent-optimal regret for markets satisfying {\em uniqueness consistency} – ma
rkets where leaving participants don’t alter the original stable matching. Previ
ously, $\Theta(\log(T))$ regret was achievable (Sankararaman et al.\,2020, Liu e
t al.\,2020b) in the much restricted {\em serial dictatorship} setting, when all
 arms have the same preference over the agents. We propose a phase based algorit
hm, where in each phase, besides deleting the globally communicated dominated ar
ms the agents locally delete arms with which they collide often. This \emph{loca
l deletion} is pivotal in breaking deadlocks arising from rank heterogeneity of
agents across arms. We further demonstrate superiority of our algorithm over exi
sting works through simulations.

****
Optimal Thompson Sampling strategies for support-aware CVaR bandits
Dorian Baudry, Romain Gautron, Emilie Kaufmann, Odalric Maillard
In this paper we study a multi-arm bandit problem in which the quality of each arm is measured by the Conditional Value at Risk (CVaR) at some level alpha of the reward distribution. While existing works in this setting mainly focus on Upper Confidence Bound algorithms, we introduce a new Thompson Sampling approach for CVaR bandits on bounded rewards that is flexible enough to solve a variety of problems grounded on physical resources. Building on a recent work by Riou & Honda (2020), we introduce B-CVTS for continuous bounded rewards and M-CVTS for multinomial distributions. On the theoretical side, we provide a non-trivial extension of their analysis that enables to theoretically bound their CVaR regret minimization performance. Strikingly, our results show that these strategies are the first to provably achieve asymptotic optimality in CVaR bandits, matching the corresponding asymptotic lower bounds for this setting. Further, we illustrate empirically the benefit of Thompson Sampling approaches both in a realistic environment simulating a use-case in agriculture and on various synthetic examples.
****
On Limited-Memory Subsampling Strategies for Bandits
Dorian Baudry, Yoan Russac, Olivier Cappé
There has been a recent surge of interest in non-parametric bandit algorithms based on subsampling. One drawback however of these approaches is the additional complexity required by random subsampling and the storage of the full history of rewards. Our first contribution is to show that a simple deterministic subsampling rule, proposed in the recent work of \citet{baudry2020sub} under the name of “last-block subsampling”, is asymptotically optimal in one-parameter exponential families. In addition, we prove that these guarantees also hold when limiting the algorithm memory to a polylogarithmic function of the time horizon. These findings open up new perspectives, in particular for non-stationary scenarios in which the arm distributions evolve over time. We propose a variant of the algorithm in which only the most recent observations are used for subsampling, achieving optimal regret guarantees under the assumption of a known number of abrupt changes. Extensive numerical simulations highlight the merits of this approach, particularly when the changes are not only affecting the means of the rewards.
****
Generalized Doubly Reparameterized Gradient Estimators
Matthias Bauer, Andriy Mnih
Efficient low-variance gradient estimation enabled by the reparameterization trick (RT) has been essential to the success of variational autoencoders. Doubly-reparameterized gradients (DReGs) improve on the RT for multi-sample variational bounds by applying reparameterization a second time for an additional reduction in variance. Here, we develop two generalizations of the DReGs estimator and show that they can be used to train conditional and hierarchical VAEs on image modelling tasks more effectively. We first extend the estimator to hierarchical models with several stochastic layers by showing how to treat additional score function terms due to the hierarchical variational posterior. We then generalize DReGs to score functions of arbitrary distributions instead of just those of the sampling distribution, which makes the estimator applicable to the parameters of the prior in addition to those of the posterior.
****
Directional Graph Networks
Dominique Beaini, Saro Passaro, Vincent Létourneau, Will Hamilton, Gabriele Corso, Pietro Lió
The lack of anisotropic kernels in graph neural networks (GNNs) strongly limits their expressiveness, contributing to well-known issues such as over-smoothing. To overcome this limitation, we propose the first globally consistent anisotropic kernels for GNNs, allowing for graph convolutions that are defined according to topologicaly-derived directional flows. First, by defining a vector field in the graph, we develop a method of applying directional derivatives and smoothing by projecting node-specific messages into the field. Then, we propose the use of

the Laplacian eigenvectors as such vector field. We show that the method generalizes CNNs on an n-dimensional grid and is provably more discriminative than standard GNNs regarding the Weisfeiler-Lehman 1-WL test. We evaluate our method on different standard benchmarks and see a relative error reduction of 8% on the CIFAR10 graph dataset and 11% to 32% on the molecular ZINC dataset, and a relative increase in precision of 1.6% on the MolPCBA dataset. An important outcome of this work is that it enables graph networks to embed directions in an unsupervised way, thus allowing a better representation of the anisotropic features in different physical or biological problems.
****
Policy Analysis using Synthetic Controls in Continuous-Time
Alexis Bellot, Mihaela van der Schaar
Counterfactual estimation using synthetic controls is one of the most successful recent methodological developments in causal inference. Despite its popularity, the current description only considers time series aligned across units and synthetic controls expressed as linear combinations of observed control units. We propose a continuous-time alternative that models the latent counterfactual path explicitly using the formalism of controlled differential equations. This model is directly applicable to the general setting of irregularly-aligned multivariate time series and may be optimized in rich function spaces – thereby improving on some limitations of existing approaches.
****
Loss Surface Simplexes for Mode Connecting Volumes and Fast Ensembling
Gregory Benton, Wesley Maddox, Sanae Lotfi, Andrew Gordon Gordon Wilson
With a better understanding of the loss surfaces for multilayer networks, we can build more robust and accurate training procedures. Recently it was discovered that independently trained SGD solutions can be connected along one-dimensional paths of near-constant training loss. In this paper, we in fact demonstrate the existence of mode-connecting simplicial complexes that form multi-dimensional manifolds of low loss, connecting many independently trained models. Building on this discovery, we show how to efficiently construct simplicial complexes for fast ensembling, outperforming independently trained deep ensembles in accuracy, calibration, and robustness to dataset shift. Notably, our approach is easy to apply and only requires a few training epochs to discover a low-loss simplex.
****
TFix: Learning to Fix Coding Errors with a Text-to-Text Transformer
Berkay Berabi, Jingxuan He, Veselin Raychev, Martin Vechev
The problem of fixing errors in programs has attracted substantial interest over the years. The key challenge for building an effective code fixing tool is to capture a wide range of errors and meanwhile maintain high accuracy. In this paper, we address this challenge and present a new learning-based system, called TFix. TFix works directly on program text and phrases the problem of code fixing as a text-to-text task. In turn, this enables it to leverage a powerful Transformer based model pre-trained on natural language and fine-tuned to generate code fixes (via a large, high-quality dataset obtained from GitHub commits). TFix is not specific to a particular programming language or class of defects and, in fact, improved its precision by simultaneously fine-tuning on 52 different error types reported by a popular static analyzer. Our evaluation on a massive dataset of JavaScript programs shows that TFix is practically effective: it is able to synthesize code that fixes the error in 67 percent of cases and significantly outperforms existing learning-based approaches.
****
Learning Queueing Policies for Organ Transplantation Allocation using Interpretable Counterfactual Survival Analysis
Jeroen Berrevoets, Ahmed Alaa, Zhaozhi Qian, James Jordon, Alexander E. S. Gimson, Mihaela van der Schaar
Organ transplantation is often the last resort for treating end-stage illnesses, but managing transplant wait-lists is challenging because of organ scarcity and the complexity of assessing donor-recipient compatibility. In this paper, we develop a data-driven model for (real-time) organ allocation using observational d

ata for transplant outcomes. Our model integrates a queuing-theoretic framework
with unsupervised learning to cluster the organs into “organ types”, and then co
nstruct priority queues (associated with each organ type) wherein incoming patie
nts are assigned. To reason about organ allocations, the model uses synthetic co
ntrols to infer a patient’s survival outcomes under counterfactual allocations t
o the different organ types{–} the model is trained end-to-end to optimise the t
rade-off between patient waiting time and expected survival time. The usage of s
ynthetic controls enable patient-level interpretations of allocation decisions t
hat can be presented and understood by clinicians. We test our model on multiple
 data sets, and show that it outperforms other organ-allocation policies in term
s of added life-years, and death count. Furthermore, we introduce a novel organ-
allocation simulator to accurately test new policies.
****
Learning from Biased Data: A Semi-Parametric Approach
Patrice Bertail, Stephan Clémençon, Yannick Guyonvarch, Nathan Noiry
We consider risk minimization problems where the (source) distribution P_S of
the training observations $Z_1, \ldots, Z_n$ differs from the (target) distribut
ion P_T involved in the risk that one seeks to minimize. Under the natural ass
umption that P_S dominates P_T, \textit{i.e.} $P_T< \! \!
Cite this Paper

 BibTeX

```
@InProceedings{pmlr-v139-bertail21a,
  title = ■ {Learning from Biased Data: A Semi-Parametric Approach},
  author =       {Bertail, Patrice and Cl{\'e}men{\c{c}}on, Stephan and Guyonvar
ch, Yannick and Noiry, Nathan},
  booktitle = ■ {Proceedings of the 38th International Conference on Machine Lea
rning},
  pages = ■ {803--812},
  year = ■ {2021},
  editor = ■ {Meila, Marina and Zhang, Tong},
  volume = ■ {139},
  series = ■ {Proceedings of Machine Learning Research},
  month = ■ {18--24 Jul},
  publisher =    {PMLR},
  pdf = ■ {http://proceedings.mlr.press/v139/bertail21a/bertail21a.pdf},
  url = ■ {https://proceedings.mlr.press/v139/bertail21a.html},
  abstract = ■ {We consider risk minimization problems where the (source) distri
bution $P_S$ of the training observations $Z_1, \ldots, Z_n$ differs from the (t
arget) distribution $P_T$ involved in the risk that one seeks to minimize. Under
 the natural assumption that $P_S$ dominates $P_T$, \textit{i.e.} $P_T< \! \!
```
Copy to ClipboardDownload
 Endnote
```
%0 Conference Paper
%T Learning from Biased Data: A Semi-Parametric Approach
%A Patrice Bertail
%A Stephan Clémençon
%A Yannick Guyonvarch
%A Nathan Noiry
%B Proceedings of the 38th International Conference on Machine Learning
%C Proceedings of Machine Learning Research
%D 2021
%E Marina Meila
%E Tong Zhang■
%F pmlr-v139-bertail21a
%I PMLR
%P 803--812
%U https://proceedings.mlr.press/v139/bertail21a.html
%V 139
```

%X We consider risk minimization problems where the (source) distribution P_S of the training observations $Z_1, \ldots, Z_n$ differs from the (target) distribution P_T involved in the risk that one seeks to minimize. Under the natural assumption that P_S dominates P_T, \textit{i.e.} $P_T< \! \!
Copy to ClipboardDownload
APA

Bertail, P., Clémençon, S., Guyonvarch, Y. & Noiry, N.. (2021). Learning from Biased Data: A Semi-Parametric Approach. Proceedings of the 38th International Conference on Machine Learning, in Proceedings of Machine Learning Research 139:803-812 Available from https://proceedings.mlr.press/v139/bertail21a.html.

Copy to ClipboardDownload

Related Material

Download PDF
Supplementary ZIP
****
Is Space-Time Attention All You Need for Video Understanding?
Gedas Bertasius, Heng Wang, Lorenzo Torresani
We present a convolution-free approach to video classification built exclusively on self-attention over space and time. Our method, named “TimeSformer,” adapts the standard Transformer architecture to video by enabling spatiotemporal feature learning directly from a sequence of frame-level patches. Our experimental study compares different self-attention schemes and suggests that “divided attention,” where temporal attention and spatial attention are separately applied within each block, leads to the best video classification accuracy among the design choices considered. Despite the radically new design, TimeSformer achieves state-of-the-art results on several action recognition benchmarks, including the best reported accuracy on Kinetics-400 and Kinetics-600. Finally, compared to 3D convolutional networks, our model is faster to train, it can achieve dramatically higher test efficiency (at a small drop in accuracy), and it can also be applied to much longer video clips (over one minute long). Code and models are available at: https://github.com/facebookresearch/TimeSformer.
****
Confidence Scores Make Instance-dependent Label-noise Learning Possible
Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, Masashi Sugiyama
In learning with noisy labels, for every instance, its label can randomly walk to other classes following a transition distribution which is named a noise model. Well-studied noise models are all instance-independent, namely, the transition depends only on the original label but not the instance itself, and thus they are less practical in the wild. Fortunately, methods based on instance-dependent noise have been studied, but most of them have to rely on strong assumptions on the noise models. To alleviate this issue, we introduce confidence-scored instance-dependent noise (CSIDN), where each instance-label pair is equipped with a confidence score. We find that with the help of confidence scores, the transition distribution of each instance can be approximately estimated. Similarly to the powerful forward correction for instance-independent noise, we propose a novel instance-level forward correction for CSIDN. We demonstrate the utility and effectiveness of our method through multiple experiments on datasets with synthetic label noise and real-world unknown noise.
****
Size-Invariant Graph Representations for Graph Classification Extrapolations
Beatrice Bevilacqua, Yangze Zhou, Bruno Ribeiro
In general, graph representation learning methods assume that the train and test data come from the same distribution. In this work we consider an underexplored area of an otherwise rapidly developing field of graph representation learning:

The task of out-of-distribution (OOD) graph classification, where train and test data have different distributions, with test data unavailable during training. Our work shows it is possible to use a causal model to learn approximately invariant representations that better extrapolate between train and test data. Finally, we conclude with synthetic and real-world dataset experiments showcasing the benefits of representations that are invariant to train/test distribution shifts.
****
Principal Bit Analysis: Autoencoding with Schur-Concave Loss
Sourbh Bhadane, Aaron B Wagner, Jayadev Acharya
We consider a linear autoencoder in which the latent variables are quantized, or corrupted by noise, and the constraint is Schur-concave in the set of latent variances. Although finding the optimal encoder/decoder pair for this setup is a nonconvex optimization problem, we show that decomposing the source into its principal components is optimal. If the constraint is strictly Schur-concave and the empirical covariance matrix has only simple eigenvalues, then any optimal encoder/decoder must decompose the source in this way. As one application, we consider a strictly Schur-concave constraint that estimates the number of bits needed to represent the latent variables under fixed-rate encoding, a setup that we call \emph{Principal Bit Analysis (PBA)}. This yields a practical, general-purpose, fixed-rate compressor that outperforms existing algorithms. As a second application, we show that a prototypical autoencoder-based variable-rate compressor is guaranteed to decompose the source into its principal components.
****
Lower Bounds on Cross-Entropy Loss in the Presence of Test-time Adversaries
Arjun Nitin Bhagoji, Daniel Cullina, Vikash Sehwag, Prateek Mittal
Understanding the fundamental limits of robust supervised learning has emerged as a problem of immense interest, from both practical and theoretical standpoints. In particular, it is critical to determine classifier-agnostic bounds on the training loss to establish when learning is possible. In this paper, we determine optimal lower bounds on the cross-entropy loss in the presence of test-time adversaries, along with the corresponding optimal classification outputs. Our formulation of the bound as a solution to an optimization problem is general enough to encompass any loss function depending on soft classifier outputs. We also propose and provide a proof of correctness for a bespoke algorithm to compute this lower bound efficiently, allowing us to determine lower bounds for multiple practical datasets of interest. We use our lower bounds as a diagnostic tool to determine the effectiveness of current robust training methods and find a gap from optimality at larger budgets. Finally, we investigate the possibility of using of optimal classification outputs as soft labels to empirically improve robust training.
****
Additive Error Guarantees for Weighted Low Rank Approximation
Aditya Bhaskara, Aravinda Kanchana Ruwanpathirana, Maheshakya Wijewardena
Low-rank approximation is a classic tool in data analysis, where the goal is to approximate a matrix A with a low-rank matrix L so as to minimize the error $\norm{A - L}_F^2$. However in many applications, approximating some entries is more important than others, which leads to the weighted low rank approximation problem. However, the addition of weights makes the low-rank approximation problem intractable. Thus many works have obtained efficient algorithms under additional structural assumptions on the weight matrix (such as low rank, and appropriate block structure). We study a natural greedy algorithm for weighted low rank approximation and develop a simple condition under which it yields bi-criteria approximation up to a small additive factor in the error. The algorithm involves iteratively computing the top singular vector of an appropriately varying matrix, and is thus easy to implement at scale. Our methods also allow us to study the problem of low rank approximation under ℓ_p norm error.
****
Sample Complexity of Robust Linear Classification on Separated Data
Robi Bhattacharjee, Somesh Jha, Kamalika Chaudhuri

We consider the sample complexity of learning with adversarial robustness. Most
prior theoretical results for this problem have considered a setting where diffe
rent classes in the data are close together or overlapping. We consider, in cont
rast, the well-separated case where there exists a classifier with perfect accur
acy and robustness, and show that the sample complexity narrates an entirely dif
ferent story. Specifically, for linear classifiers, we show a large class of wel
l-separated distributions where the expected robust loss of any algorithm is at
least $\Omega(\frac{d}{n})$, whereas the max margin algorithm has expected stand
ard loss $O(\frac{1}{n})$. This shows a gap in the standard and robust losses th
at cannot be obtained via prior techniques. Additionally, we present an algorith
m that, given an instance where the robustness radius is much smaller than the g
ap between the classes, gives a solution with expected robust loss is $O(\frac{1
}{n})$. This shows that for very well-separated data, convergence rates of $O(\f
rac{1}{n})$ are achievable, which is not the case otherwise. Our results apply t
o robustness measured in any ℓ_p norm with $p > 1$ (including $p = \infty$)
.
****
Finding k in Latent $k-$ polytope
Chiranjib Bhattacharyya, Ravindran Kannan, Amit Kumar
The recently introduced Latent $k-$ Polytope($\LkP$) encompasses several stochas
tic Mixed Membership models including Topic Models. The problem of finding k,
the number of extreme points of $\LkP$, is a fundamental challenge and includes
several important open problems such as determination of number of components in
 Ad-mixtures. This paper addresses this challenge by introducing Interpolative C
onvex Rank(\INR) of a matrix defined as the minimum number of its columns whose
convex hull is within Hausdorff distance ε of the convex hull of all
 columns. The first important contribution of this paper is to show that under \
emph{standard assumptions} k equals the \INR of a \emph{subset smoothed data m
atrix} defined from Data generated from an $\LkP$. The second important contribu
tion of the paper is a polynomial time algorithm for finding k under standard
assumptions. An immediate corollary is the first polynomial time algorithm for f
inding the \emph{inner dimension} in Non-negative matrix factorisation(NMF) with
 assumptions which are qualitatively different than existing ones such as \emph{
Separability}. %An immediate corollary is the first polynomial time algorithm fo
r finding the \emph{inner dimension} in Non-negative matrix factorisation(NMF) w
ith assumptions considerably weaker than \emph{Separability}.
****
Non-Autoregressive Electron Redistribution Modeling for Reaction Prediction
Hangrui Bi, Hengyi Wang, Chence Shi, Connor Coley, Jian Tang, Hongyu Guo
Reliably predicting the products of chemical reactions presents a fundamental ch
allenge in synthetic chemistry. Existing machine learning approaches typically p
roduce a reaction product by sequentially forming its subparts or intermediate m
olecules. Such autoregressive methods, however, not only require a pre-defined o
rder for the incremental construction but preclude the use of parallel decoding
for efficient computation. To address these issues, we devise a non-autoregressi
ve learning paradigm that predicts reaction in one shot. Leveraging the fact tha
t chemical reactions can be described as a redistribution of electrons in molecu
les, we formulate a reaction as an arbitrary electron flow and predict it with a
 novel multi-pointer decoding network. Experiments on the USPTO-MIT dataset show
 that our approach has established a new state-of-the-art top-1 accuracy and ach
ieves at least 27 times inference speedup over the state-of-the-art methods. Als
o, our predictions are easier for chemists to interpret owing to predicting the
electron flows.
****
TempoRL: Learning When to Act
André Biedenkapp, Raghu Rajan, Frank Hutter, Marius Lindauer
Reinforcement learning is a powerful approach to learn behaviour through interac
tions with an environment. However, behaviours are usually learned in a purely r
eactive fashion, where an appropriate action is selected based on an observation
. In this form, it is challenging to learn when it is necessary to execute new d

ecisions. This makes learning inefficient especially in environments that need various degrees of fine and coarse control. To address this, we propose a proactive setting in which the agent not only selects an action in a state but also for how long to commit to that action. Our TempoRL approach introduces skip connections between states and learns a skip-policy for repeating the same action along these skips. We demonstrate the effectiveness of TempoRL on a variety of traditional and deep RL environments, showing that our approach is capable of learning successful policies up to an order of magnitude faster than vanilla Q-learning.
****
Follow-the-Regularized-Leader Routes to Chaos in Routing Games
Jakub Bielawski, Thiparat Chotibut, Fryderyk Falniowski, Grzegorz Kosiorowski, Micha■ Misiurewicz, Georgios Piliouras
We study the emergence of chaotic behavior of Follow-the-Regularized Leader (FoReL) dynamics in games. We focus on the effects of increasing the population size or the scale of costs in congestion games, and generalize recent results on unstable, chaotic behaviors in the Multiplicative Weights Update dynamics to a much larger class of FoReL dynamics. We establish that, even in simple linear non-atomic congestion games with two parallel links and \emph{any} fixed learning rate, unless the game is fully symmetric, increasing the population size or the scale of costs causes learning dynamics to becomes unstable and eventually chaotic, in the sense of Li-Yorke and positive topological entropy. Furthermore, we prove the existence of novel non-standard phenomena such as the coexistence of stable Nash equilibria and chaos in the same game. We also observe the simultaneous creation of a chaotic attractor as another chaotic attractor gets destroyed. Lastly, although FoReL dynamics can be strange and non-equilibrating, we prove that the time average still converges to an \emph{exact} equilibrium for any choice of learning rate and any scale of costs.
****
Neural Symbolic Regression that scales
Luca Biggio, Tommaso Bendinelli, Alexander Neitz, Aurelien Lucchi, Giambattista Parascandolo
Symbolic equations are at the core of scientific discovery. The task of discovering the underlying equation from a set of input-output pairs is called symbolic regression. Traditionally, symbolic regression methods use hand-designed strategies that do not improve with experience. In this paper, we introduce the first symbolic regression method that leverages large scale pre-training. We procedurally generate an unbounded set of equations, and simultaneously pre-train a Transformer to predict the symbolic equation from a corresponding set of input-output-pairs. At test time, we query the model on a new set of points and use its output to guide the search for the equation. We show empirically that this approach can re-discover a set of well-known physical equations, and that it improves over time with more data and compute.
****
Model Distillation for Revenue Optimization: Interpretable Personalized Pricing
Max Biggs, Wei Sun, Markus Ettl
Data-driven pricing strategies are becoming increasingly common, where customers are offered a personalized price based on features that are predictive of their valuation of a product. It is desirable for this pricing policy to be simple and interpretable, so it can be verified, checked for fairness, and easily implemented. However, efforts to incorporate machine learning into a pricing framework often lead to complex pricing policies that are not interpretable, resulting in slow adoption in practice. We present a novel, customized, prescriptive tree-based algorithm that distills knowledge from a complex black-box machine learning algorithm, segments customers with similar valuations and prescribes prices in such a way that maximizes revenue while maintaining interpretability. We quantify the regret of a resulting policy and demonstrate its efficacy in applications with both synthetic and real-world datasets.
****
Scalable Normalizing Flows for Permutation Invariant Densities
Marin Biloš, Stephan Günnemann

Modeling sets is an important problem in machine learning since this type of dat a can be found in many domains. A promising approach defines a family of permuta tion invariant densities with continuous normalizing flows. This allows us to ma ximize the likelihood directly and sample new realizations with ease. In this wo rk, we demonstrate how calculating the trace, a crucial step in this method, rai ses issues that occur both during training and inference, limiting its practical ity. We propose an alternative way of defining permutation equivariant transform ations that give closed form trace. This leads not only to improvements while tr aining, but also to better final performance. We demonstrate the benefits of our approach on point processes and general set modeling.
****
Online Learning for Load Balancing of Unknown Monotone Resource Allocation Games
Ilai Bistritz, Nicholas Bambos
Consider N players that each uses a mixture of K resources. Each of the players' reward functions includes a linear pricing term for each resource that is contr olled by the game manager. We assume that the game is strongly monotone, so if e ach player runs gradient descent, the dynamics converge to a unique Nash equilib rium (NE). Unfortunately, this NE can be inefficient since the total load on a g iven resource can be very high. In principle, we can control the total loads by tuning the coefficients of the pricing terms. However, finding pricing coefficie nts that balance the loads requires knowing the players' reward functions and th eir action sets. Obtaining this game structure information is infeasible in a la rge-scale network and violates the users' privacy. To overcome this, we propose a simple algorithm that learns to shift the NE of the game to meet the total loa d constraints by adjusting the pricing coefficients in an online manner. Our alg orithm only requires the total load per resource as feedback and does not need t o know the reward functions or the action sets. We prove that our algorithm guar antees convergence in L2 to a NE that meets target total load constraints. Simul ations show the effectiveness of our approach when applied to smart grid demand-side management or power control in wireless networks.
****
Low-Precision Reinforcement Learning: Running Soft Actor-Critic in Half Precisio n
Johan Björck, Xiangyu Chen, Christopher De Sa, Carla P Gomes, Kilian Weinberger
Low-precision training has become a popular approach to reduce compute requireme nts, memory footprint, and energy consumption in supervised learning. In contras t, this promising approach has not yet enjoyed similarly widespread adoption wit hin the reinforcement learning (RL) community, partly because RL agents can be n otoriously hard to train even in full precision. In this paper we consider conti nuous control with the state-of-the-art SAC agent and demonstrate that a naïve a daptation of low-precision methods from supervised learning fails. We propose a set of six modifications, all straightforward to implement, that leaves the unde rlying agent and its hyperparameters unchanged but improves the numerical stabil ity dramatically. The resulting modified SAC agent has lower memory and compute requirements while matching full-precision rewards, demonstrating that low-preci sion training can substantially accelerate state-of-the-art RL without parameter tuning.
****
Multiplying Matrices Without Multiplying
Davis Blalock, John Guttag
Multiplying matrices is among the most fundamental and most computationally dema nding operations in machine learning and scientific computing. Consequently, the task of efficiently approximating matrix products has received significant atte ntion. We introduce a learning-based algorithm for this task that greatly outper forms existing methods. Experiments using hundreds of matrices from diverse doma ins show that it often runs 10x faster than alternatives at a given level of err or, as well as 100x faster than exact matrix multiplication. In the common case that one matrix is known ahead of time, our method also has the interesting prop erty that it requires zero multiply-adds. These results suggest that a mixture o f hashing, averaging, and byte shuffling{—}the core operations of our method{—}c

ould be a more promising building block for machine learning than the sparsified, factorized, and/or scalar quantized matrix products that have recently been the focus of substantial research and hardware investment.
****
One for One, or All for All: Equilibria and Optimality of Collaboration in Federated Learning
Avrim Blum, Nika Haghtalab, Richard Lanas Phillips, Han Shao
In recent years, federated learning has been embraced as an approach for bringing about collaboration across large populations of learning agents. However, little is known about how collaboration protocols should take agents' incentives into account when allocating individual resources for communal learning in order to maintain such collaborations. Inspired by game theoretic notions, this paper introduces a framework for incentive-aware learning and data sharing in federated learning. Our stable and envy-free equilibria capture notions of collaboration in the presence of agents interested in meeting their learning objectives while keeping their own sample collection burden low. For example, in an envy-free equilibrium, no agent would wish to swap their sampling burden with any other agent and in a stable equilibrium, no agent would wish to unilaterally reduce their sampling burden. In addition to formalizing this framework, our contributions include characterizing the structural properties of such equilibria, proving when they exist, and showing how they can be computed. Furthermore, we compare the sample complexity of incentive-aware collaboration with that of optimal collaboration when one ignores agents' incentives.
****
Black-box density function estimation using recursive partitioning
Erik Bodin, Zhenwen Dai, Neill Campbell, Carl Henrik Ek
We present a novel approach to Bayesian inference and general Bayesian computation that is defined through a sequential decision loop. Our method defines a recursive partitioning of the sample space. It neither relies on gradients nor requires any problem-specific tuning, and is asymptotically exact for any density function with a bounded domain. The output is an approximation to the whole density function including the normalisation constant, via partitions organised in efficient data structures. Such approximations may be used for evidence estimation or fast posterior sampling, but also as building blocks to treat a larger class of estimation problems. The algorithm shows competitive performance to recent state-of-the-art methods on synthetic and real-world problems including parameter inference for gravitational-wave physics.
****
Weisfeiler and Lehman Go Topological: Message Passing Simplicial Networks
Cristian Bodnar, Fabrizio Frasca, Yuguang Wang, Nina Otter, Guido F Montufar, Pietro Lió, Michael Bronstein
The pairwise interaction paradigm of graph machine learning has predominantly governed the modelling of relational systems. However, graphs alone cannot capture the multi-level interactions present in many complex systems and the expressive power of such schemes was proven to be limited. To overcome these limitations, we propose Message Passing Simplicial Networks (MPSNs), a class of models that perform message passing on simplicial complexes (SCs). To theoretically analyse the expressivity of our model we introduce a Simplicial Weisfeiler-Lehman (SWL) colouring procedure for distinguishing non-isomorphic SCs. We relate the power of SWL to the problem of distinguishing non-isomorphic graphs and show that SWL and MPSNs are strictly more powerful than the WL test and not less powerful than the 3-WL test. We deepen the analysis by comparing our model with traditional graph neural networks (GNNs) with ReLU activations in terms of the number of linear regions of the functions they can represent. We empirically support our theoretical claims by showing that MPSNs can distinguish challenging strongly regular graphs for which GNNs fail and, when equipped with orientation equivariant layers, they can improve classification accuracy in oriented SCs compared to a GNN baseline.
****
The Hintons in your Neural Network: a Quantum Field Theory View of Deep Learning

Roberto Bondesan, Max Welling

In this work we develop a quantum field theory formalism for deep learning, where input signals are encoded in Gaussian states, a generalization of Gaussian processes which encode the agent’s uncertainty about the input signal. We show how to represent linear and non-linear layers as unitary quantum gates, and interpret the fundamental excitations of the quantum model as particles, dubbed “Hintons”. On top of opening a new perspective and techniques for studying neural networks, the quantum formulation is well suited for optical quantum computing, and provides quantum deformations of neural networks that can be run efficiently on those devices. Finally, we discuss a semi-classical limit of the quantum deformed models which is amenable to classical simulation.

****

Offline Contextual Bandits with Overparameterized Models

David Brandfonbrener, William Whitney, Rajesh Ranganath, Joan Bruna

Recent results in supervised learning suggest that while overparameterized models have the capacity to overfit, they in fact generalize quite well. We ask whether the same phenomenon occurs for offline contextual bandits. Our results are mixed. Value-based algorithms benefit from the same generalization behavior as overparameterized supervised learning, but policy-based algorithms do not. We show that this discrepancy is due to the \emph{action-stability} of their objectives. An objective is action-stable if there exists a prediction (action-value vector or action distribution) which is optimal no matter which action is observed. While value-based objectives are action-stable, policy-based objectives are unstable. We formally prove upper bounds on the regret of overparameterized value-based learning and lower bounds on the regret for policy-based algorithms. In our experiments with large neural networks, this gap between action-stable value-based objectives and unstable policy-based objectives leads to significant performance differences.

****

High-Performance Large-Scale Image Recognition Without Normalization

Andy Brock, Soham De, Samuel L Smith, Karen Simonyan

Batch normalization is a key component of most image classification models, but it has many undesirable properties stemming from its dependence on the batch size and interactions between examples. Although recent work has succeeded in training deep ResNets without normalization layers, these models do not match the test accuracies of the best batch-normalized networks, and are often unstable for large learning rates or strong data augmentations. In this work, we develop an adaptive gradient clipping technique which overcomes these instabilities, and design a significantly improved class of Normalizer-Free ResNets. Our smaller models match the test accuracy of an EfficientNet-B7 on ImageNet while being up to 8.7x faster to train, and our largest models attain a new state-of-the-art top-1 accuracy of 86.5%. In addition, Normalizer-Free models attain significantly better performance than their batch-normalized counterparts when fine-tuning on ImageNet after large-scale pre-training on a dataset of 300 million labeled images, with our best models obtaining an accuracy of 89.2%.

****

Evaluating the Implicit Midpoint Integrator for Riemannian Hamiltonian Monte Carlo

James Brofos, Roy R Lederman

Riemannian manifold Hamiltonian Monte Carlo is traditionally carried out using the generalized leapfrog integrator. However, this integrator is not the only choice and other integrators yielding valid Markov chain transition operators may be considered. In this work, we examine the implicit midpoint integrator as an alternative to the generalized leapfrog integrator. We discuss advantages and disadvantages of the implicit midpoint integrator for Hamiltonian Monte Carlo, its theoretical properties, and an empirical assessment of the critical attributes of such an integrator for Hamiltonian Monte Carlo: energy conservation, volume preservation, and reversibility. Empirically, we find that while leapfrog iterations are faster, the implicit midpoint integrator has better energy conservation, leading to higher acceptance rates, as well as better conservation of volume and

better reversibility, arguably yielding a more accurate sampling procedure.
****
Reinforcement Learning of Implicit and Explicit Control Flow Instructions
Ethan Brooks, Janarthanan Rajendran, Richard L Lewis, Satinder Singh
Learning to flexibly follow task instructions in dynamic environments poses interesting challenges for reinforcement learning agents. We focus here on the problem of learning control flow that deviates from a strict step-by-step execution of instructions{—}that is, control flow that may skip forward over parts of the instructions or return backward to previously completed or skipped steps. Demand for such flexible control arises in two fundamental ways: explicitly when control is specified in the instructions themselves (such as conditional branching and looping) and implicitly when stochastic environment dynamics require re-completion of instructions whose effects have been perturbed, or opportunistic skipping of instructions whose effects are already present. We formulate an attention-based architecture that meets these challenges by learning, from task reward only, to flexibly attend to and condition behavior on an internal encoding of the instructions. We test the architecture's ability to learn both explicit and implicit control in two illustrative domains—one inspired by Minecraft and the other by StarCraft—and show that the architecture exhibits zero-shot generalization to novel instructions of length greater than those in a training set, at a performance level unmatched by three baseline recurrent architectures and one ablation architecture.
****
Machine Unlearning for Random Forests
Jonathan Brophy, Daniel Lowd
Responding to user data deletion requests, removing noisy examples, or deleting corrupted training data are just a few reasons for wanting to delete instances from a machine learning (ML) model. However, efficiently removing this data from an ML model is generally difficult. In this paper, we introduce data removal-enabled (DaRE) forests, a variant of random forests that enables the removal of training data with minimal retraining. Model updates for each DaRE tree in the forest are exact, meaning that removing instances from a DaRE model yields exactly the same model as retraining from scratch on updated data. DaRE trees use randomness and caching to make data deletion efficient. The upper levels of DaRE trees use random nodes, which choose split attributes and thresholds uniformly at random. These nodes rarely require updates because they only minimally depend on the data. At the lower levels, splits are chosen to greedily optimize a split criterion such as Gini index or mutual information. DaRE trees cache statistics at each node and training data at each leaf, so that only the necessary subtrees are updated as data is removed. For numerical attributes, greedy nodes optimize over a random subset of thresholds, so that they can maintain statistics while approximating the optimal threshold. By adjusting the number of thresholds considered for greedy nodes, and the number of random nodes, DaRE trees can trade off between more accurate predictions and more efficient updates. In experiments on 13 real-world datasets and one synthetic dataset, we find DaRE forests delete data orders of magnitude faster than retraining from scratch while sacrificing little to no predictive power.
****
Value Alignment Verification
Daniel S Brown, Jordan Schneider, Anca Dragan, Scott Niekum
As humans interact with autonomous agents to perform increasingly complicated, potentially risky tasks, it is important to be able to efficiently evaluate an agent's performance and correctness. In this paper we formalize and theoretically analyze the problem of efficient value alignment verification: how to efficiently test whether the behavior of another agent is aligned with a human's values? The goal is to construct a kind of "driver's test" that a human can give to any agent which will verify value alignment via a minimal number of queries. We study alignment verification problems with both idealized humans that have an explicit reward function as well as problems where they have implicit values. We analyze verification of exact value alignment for rational agents, propose and test he

uristics for value alignment verification in gridworlds and a continuous autonomous driving domain, and prove that there exist sufficient conditions such that we can verify epsilon-alignment in any environment via a constant-query-complexity alignment test.
****
Model-Free and Model-Based Policy Evaluation when Causality is Uncertain
David A Bruns-Smith
When decision-makers can directly intervene, policy evaluation algorithms give valid causal estimates. In off-policy evaluation (OPE), there may exist unobserved variables that both impact the dynamics and are used by the unknown behavior policy. These “confounders” will introduce spurious correlations and naive estimates for a new policy will be biased. We develop worst-case bounds to assess sensitivity to these unobserved confounders in finite horizons when confounders are drawn iid each period. We demonstrate that a model-based approach with robust MDPs gives sharper lower bounds by exploiting domain knowledge about the dynamics. Finally, we show that when unobserved confounders are persistent over time, OPE is far more difficult and existing techniques produce extremely conservative bounds.
****
Narrow Margins: Classification, Margins and Fat Tails
Francois Buet-Golfouse
It is well-known that, for separable data, the regularised two-class logistic regression or support vector machine re-normalised estimate converges to the maximal margin classifier as the regularisation hyper-parameter λ goes to 0. The fact that different loss functions may lead to the same solution is of theoretical and practical relevance as margin maximisation allows more straightforward considerations in terms of generalisation and geometric interpretation. We investigate the case where this convergence property is not guaranteed to hold and show that it can be fully characterised by the distribution of error terms in the latent variable interpretation of linear classifiers. In particular, if errors follow a regularly varying distribution, then the regularised and re-normalised estimate does not converge to the maximal margin classifier. This shows that classification with fat tails has a qualitatively different behaviour, which should be taken into account when considering real-life data.
****
Differentially Private Correlation Clustering
Mark Bun, Marek Elias, Janardhan Kulkarni
Correlation clustering is a widely used technique in unsupervised machine learning. Motivated by applications where individual privacy is a concern, we initiate the study of differentially private correlation clustering. We propose an algorithm that achieves subquadratic additive error compared to the optimal cost. In contrast, straightforward adaptations of existing non-private algorithms all lead to a trivial quadratic error. Finally, we give a lower bound showing that any pure differentially private algorithm for correlation clustering requires additive error Ω(n).
****
Disambiguation of Weak Supervision leading to Exponential Convergence rates
Vivien A Cabannnes, Francis Bach, Alessandro Rudi
Machine learning approached through supervised learning requires expensive annotation of data. This motivates weakly supervised learning, where data are annotated with incomplete yet discriminative information. In this paper, we focus on partial labelling, an instance of weak supervision where, from a given input, we are given a set of potential targets. We review a disambiguation principle to recover full supervision from weak supervision, and propose an empirical disambiguation algorithm. We prove exponential convergence rates of our algorithm under classical learnability assumptions, and we illustrate the usefulness of our method on practical examples.
****
Finite mixture models do not reliably learn the number of components
Diana Cai, Trevor Campbell, Tamara Broderick

Scientists and engineers are often interested in learning the number of subpopul ations (or components) present in a data set. A common suggestion is to use a fi nite mixture model (FMM) with a prior on the number of components. Past work has shown the resulting FMM component-count posterior is consistent; that is, the p osterior concentrates on the true, generating number of components. But consiste ncy requires the assumption that the component likelihoods are perfectly specifi ed, which is unrealistic in practice. In this paper, we add rigor to data-analys is folk wisdom by proving that under even the slightest model misspecification, the FMM component-count posterior diverges: the posterior probability of any par ticular finite number of components converges to 0 in the limit of infinite data . Contrary to intuition, posterior-density consistency is not sufficient to esta blish this result. We develop novel sufficient conditions that are more realisti c and easily checkable than those common in the asymptotics literature. We illus trate practical consequences of our theory on simulated and real data.
****

A Theory of Label Propagation for Subpopulation Shift

Tianle Cai, Ruiqi Gao, Jason Lee, Qi Lei

One of the central problems in machine learning is domain adaptation. Different from past theoretical works, we consider a new model of subpopulation shift in t he input or representation space. In this work, we propose a provably effective framework based on label propagation by using an input consistency loss. In our analysis we used a simple but realistic “expansion” assumption, which has been p roposed in \citet{wei2021theoretical}. It turns out that based on a teacher clas sifier on the source domain, the learned classifier can not only propagate to th e target domain but also improve upon the teacher. By leveraging existing genera lization bounds, we also obtain end-to-end finite-sample guarantees on deep neur al networks. In addition, we extend our theoretical framework to a more general setting of source-to-target transfer based on an additional unlabeled dataset, w hich can be easily applied to various learning scenarios. Inspired by our theory , we adapt consistency-based semi-supervised learning methods to domain adaptati on settings and gain significant improvements.
****

Lenient Regret and Good-Action Identification in Gaussian Process Bandits

Xu Cai, Selwyn Gomes, Jonathan Scarlett

In this paper, we study the problem of Gaussian process (GP) bandits under relax ed optimization criteria stating that any function value above a certain thresho ld is “good enough”. On the theoretical side, we study various {\em lenient regr et} notions in which all near-optimal actions incur zero penalty, and provide up per bounds on the lenient regret for GP-UCB and an elimination algorithm, circum venting the usual $O(\sqrt{T})$ term (with time horizon T) resulting from zoom ing extremely close towards the function maximum. In addition, we complement the se upper bounds with algorithm-independent lower bounds. On the practical side, we consider the problem of finding a single “good action” according to a known p re-specified threshold, and introduce several good-action identification algorit hms that exploit knowledge of the threshold. We experimentally find that such al gorithms can typically find a good action faster than standard optimization-base d approaches.
****

A Zeroth-Order Block Coordinate Descent Algorithm for Huge-Scale Black-Box Optim ization

Hanqin Cai, Yuchen Lou, Daniel Mckenzie, Wotao Yin

We consider the zeroth-order optimization problem in the huge-scale setting, whe re the dimension of the problem is so large that performing even basic vector op erations on the decision variables is infeasible. In this paper, we propose a no vel algorithm, coined ZO-BCD, that exhibits favorable overall query complexity a nd has a much smaller per-iteration computational complexity. In addition, we di scuss how the memory footprint of ZO-BCD can be reduced even further by the clev er use of circulant measurement matrices. As an application of our new method, w e propose the idea of crafting adversarial attacks on neural network based class ifiers in a wavelet domain, which can result in problem dimensions of over one m

illion. In particular, we show that crafting adversarial examples to audio classifiers in a wavelet domain can achieve the state-of-the-art attack success rate of 97.9% with significantly less distortion.
****
GraphNorm: A Principled Approach to Accelerating Graph Neural Network Training
Tianle Cai, Shengjie Luo, Keyulu Xu, Di He, Tie-Yan Liu, Liwei Wang
Normalization is known to help the optimization of deep neural networks. Curiously, different architectures require specialized normalization methods. In this paper, we study what normalization is effective for Graph Neural Networks (GNNs). First, we adapt and evaluate the existing methods from other domains to GNNs. Faster convergence is achieved with InstanceNorm compared to BatchNorm and LayerNorm. We provide an explanation by showing that InstanceNorm serves as a preconditioner for GNNs, but such preconditioning effect is weaker with BatchNorm due to the heavy batch noise in graph datasets. Second, we show that the shift operation in InstanceNorm results in an expressiveness degradation of GNNs for highly regular graphs. We address this issue by proposing GraphNorm with a learnable shift. Empirically, GNNs with GraphNorm converge faster compared to GNNs using other normalization. GraphNorm also improves the generalization of GNNs, achieving better performance on graph classification benchmarks.
****
On Lower Bounds for Standard and Robust Gaussian Process Bandit Optimization
Xu Cai, Jonathan Scarlett
In this paper, we consider algorithm independent lower bounds for the problem of black-box optimization of functions having a bounded norm is some Reproducing Kernel Hilbert Space (RKHS), which can be viewed as a non-Bayesian Gaussian process bandit problem. In the standard noisy setting, we provide a novel proof technique for deriving lower bounds on the regret, with benefits including simplicity, versatility, and an improved dependence on the error probability. In a robust setting in which the final point is perturbed by an adversary, we strengthen an existing lower bound that only holds for target success probabilities very close to one, by allowing for arbitrary target success probabilities in (0, 1). Furthermore, in a distinct robust setting in which every sampled point may be perturbed by a constrained adversary, we provide a novel lower bound for deterministic strategies, demonstrating an inevitable joint dependence of the cumulative regret on the corruption level and the time horizon, in contrast with existing lower bounds that only characterize the individual dependencies.
****
High-dimensional Experimental Design and Kernel Bandits
Romain Camilleri, Kevin Jamieson, Julian Katz-Samuels
In recent years methods from optimal linear experimental design have been leveraged to obtain state of the art results for linear bandits. A design returned from an objective such as G-optimal design is actually a probability distribution over a pool of potential measurement vectors. Consequently, one nuisance of the approach is the task of converting this continuous probability distribution into a discrete assignment of N measurements. While sophisticated rounding techniques have been proposed, in d dimensions they require N to be at least d, d log(log(d)), or d^2 based on the sub-optimality of the solution. In this paper we are interested in settings where N may be much less than d, such as in experimental design in an RKHS where d may be effectively infinite. In this work, we propose a rounding procedure that frees N of any dependence on the dimension d, while achieving nearly the same performance guarantees of existing rounding procedures. We evaluate the procedure against a baseline that projects the problem to a lower dimensional space and performs rounding there, which requires N to just be at least a notion of the effective dimension. We also leverage our new approach in a new algorithm for kernelized bandits to obtain state of the art results for regret minimization and pure exploration. An advantage of our approach over existing UCB-like approaches is that our kernel bandit algorithms are provably robust to model misspecification.
****
A Gradient Based Strategy for Hamiltonian Monte Carlo Hyperparameter Optimizatio

n
Andrew Campbell, Wenlong Chen, Vincent Stimper, Jose Miguel Hernandez-Lobato, Yi chuan Zhang
Hamiltonian Monte Carlo (HMC) is one of the most successful sampling methods in machine learning. However, its performance is significantly affected by the choi ce of hyperparameter values. Existing approaches for optimizing the HMC hyperpar ameters either optimize a proxy for mixing speed or consider the HMC chain as an implicit variational distribution and optimize a tractable lower bound that can be very loose in practice. Instead, we propose to optimize an objective that qu antifies directly the speed of convergence to the target distribution. Our objec tive can be easily optimized using stochastic gradient descent. We evaluate our proposed method and compare to baselines on a variety of problems including samp ling from synthetic 2D distributions, reconstructing sparse signals, learning de ep latent variable models and sampling molecular configurations from the Boltzma nn distribution of a 22 atom molecule. We find that our method is competitive wi th or improves upon alternative baselines in all these experiments.
****
Asymmetric Heavy Tails and Implicit Bias in Gaussian Noise Injections
Alexander Camuto, Xiaoyu Wang, Lingjiong Zhu, Chris Holmes, Mert Gurbuzbalaban, Umut Simsekli
Gaussian noise injections (GNIs) are a family of simple and widely-used regulari sation methods for training neural networks, where one injects additive or multi plicative Gaussian noise to the network activations at every iteration of the op timisation algorithm, which is typically chosen as stochastic gradient descent (SGD). In this paper, we focus on the so-called ‘implicit effect’ of GNIs, which is the effect of the injected noise on the dynamics of SGD. We show that this ef fect induces an \emph{asymmetric heavy-tailed noise} on SGD gradient updates. In order to model this modified dynamics, we first develop a Langevin-like stochas tic differential equation that is driven by a general family of \emph{asymmetric } heavy-tailed noise. Using this model we then formally prove that GNIs induce a n ‘implicit bias’, which varies depending on the heaviness of the tails and the level of asymmetry. Our empirical results confirm that different types of neural networks trained with GNIs are well-modelled by the proposed dynamics and that the implicit effect of these injections induces a bias that degrades the perform ance of networks.
****
Fold2Seq: A Joint Sequence(1D)-Fold(3D) Embedding-based Generative Model for Pro tein Design
Yue Cao, Payel Das, Vijil Chenthamarakshan, Pin-Yu Chen, Igor Melnyk, Yang Shen
Designing novel protein sequences for a desired 3D topological fold is a fundame ntal yet non-trivial task in protein engineering. Challenges exist due to the co mplex sequence–fold relationship, as well as the difficulties to capture the div ersity of the sequences (therefore structures and functions) within a fold. To o vercome these challenges, we propose Fold2Seq, a novel transformer-based generat ive framework for designing protein sequences conditioned on a specific target f old. To model the complex sequence–structure relationship, Fold2Seq jointly lear ns a sequence embedding using a transformer and a fold embedding from the densit y of secondary structural elements in 3D voxels. On test sets with single, high- resolution and complete structure inputs for individual folds, our experiments d emonstrate improved or comparable performance of Fold2Seq in terms of speed, cov erage, and reliability for sequence design, when compared to existing state-of-t he-art methods that include data-driven deep generative models and physics-based RosettaDesign. The unique advantages of fold-based Fold2Seq, in comparison to a structure-based deep model and RosettaDesign, become more evident on three addi tional real-world challenges originating from low-quality, incomplete, or ambigu ous input structures. Source code and data are available at https://github.com/I BM/fold2seq.
****
Learning from Similarity-Confidence Data
Yuzhou Cao, Lei Feng, Yitian Xu, Bo An, Gang Niu, Masashi Sugiyama

Weakly supervised learning has drawn considerable attention recently to reduce t
he expensive time and labor consumption of labeling massive data. In this paper,
 we investigate a novel weakly supervised learning problem of learning from simi
larity-confidence (Sconf) data, where only unlabeled data pairs equipped with co
nfidence that illustrates their degree of similarity (two examples are similar i
f they belong to the same class) are needed for training a discriminative binary
 classifier. We propose an unbiased estimator of the classification risk that ca
n be calculated from only Sconf data and show that the estimation error bound ac
hieves the optimal convergence rate. To alleviate potential overfitting when fle
xible models are used, we further employ a risk correction scheme on the propose
d risk estimator. Experimental results demonstrate the effectiveness of the prop
osed methods.
****
Parameter-free Locally Accelerated Conditional Gradients
Alejandro Carderera, Jelena Diakonikolas, Cheuk Yin Lin, Sebastian Pokutta
Projection-free conditional gradient (CG) methods are the algorithms of choice f
or constrained optimization setups in which projections are often computationall
y prohibitive but linear optimization over the constraint set remains computatio
nally feasible. Unlike in projection-based methods, globally accelerated converg
ence rates are in general unattainable for CG. However, a very recent work on Lo
cally accelerated CG (LaCG) has demonstrated that local acceleration for CG is p
ossible for many settings of interest. The main downside of LaCG is that it requ
ires knowledge of the smoothness and strong convexity parameters of the objectiv
e function. We remove this limitation by introducing a novel, Parameter-Free Loc
ally accelerated CG (PF-LaCG) algorithm, for which we provide rigorous convergen
ce guarantees. Our theoretical results are complemented by numerical experiments
, which demonstrate local acceleration and showcase the practical improvements o
f PF-LaCG over non-accelerated algorithms, both in terms of iteration count and
wall-clock time.
****
Optimizing persistent homology based functions
Mathieu Carriere, Frederic Chazal, Marc Glisse, Yuichi Ike, Hariprasad Kannan, Y
uhei Umeda
Solving optimization tasks based on functions and losses with a topological flav
or is a very active and growing field of research in data science and Topologica
l Data Analysis, with applications in non-convex optimization, statistics and ma
chine learning. However, the approaches proposed in the literature are usually a
nchored to a specific application and/or topological construction, and do not co
me with theoretical guarantees. To address this issue, we study the differentiab
ility of a general map associated with the most common topological construction,
 that is, the persistence map. Building on real analytic geometry arguments, we
propose a general framework that allows us to define and compute gradients for p
ersistence-based functions in a very simple way. We also provide a simple, expli
cit and sufficient condition for convergence of stochastic subgradient methods f
or such functions. This result encompasses all the constructions and application
s of topological optimization in the literature. Finally, we provide associated
code, that is easy to handle and to mix with other non-topological methods and c
onstraints, as well as some experiments showcasing the versatility of our approa
ch.
****
Online Policy Gradient for Model Free Learning of Linear Quadratic Regulators wi
th $\sqrt$T Regret
Asaf B Cassel, Tomer Koren
We consider the task of learning to control a linear dynamical system under fixe
d quadratic costs, known as the Linear Quadratic Regulator (LQR) problem. While
model-free approaches are often favorable in practice, thus far only model-based
 methods, which rely on costly system identification, have been shown to achieve
 regret that scales with the optimal dependence on the time horizon T. We presen
t the first model-free algorithm that achieves similar regret guarantees. Our me
thod relies on an efficient policy gradient scheme, and a novel and tighter anal

ysis of the cost of exploration in policy space in this setting.
****
Multi-Receiver Online Bayesian Persuasion
Matteo Castiglioni, Alberto Marchesi, Andrea Celli, Nicola Gatti
Bayesian persuasion studies how an informed sender should partially disclose inf
ormation to influence the behavior of a self-interested receiver. Classical mode
ls make the stringent assumption that the sender knows the receiver’s utility. T
his can be relaxed by considering an online learning framework in which the send
er repeatedly faces a receiver of an unknown, adversarially selected type. We st
udy, for the first time, an online Bayesian persuasion setting with multiple rec
eivers. We focus on the case with no externalities and binary actions, as custom
ary in offline models. Our goal is to design no-regret algorithms for the sender
 with polynomial per-iteration running time. First, we prove a negative result:
for any 0 < α $\leq$ 1, there is no polynomial-time no-α-regret al
gorithm when the sender’s utility function is supermodular or anonymous. Then, w
e focus on the setting of submodular sender’s utility functions and we show that
, in this case, it is possible to design a polynomial-time no-(1-1/e)-regret alg
orithm. To do so, we introduce a general online gradient descent framework to ha
ndle online learning problems with a finite number of possible loss functions. T
his requires the existence of an approximate projection oracle. We show that, in
 our setting, there exists one such projection oracle which can be implemented i
n polynomial time.
****
Marginal Contribution Feature Importance - an Axiomatic Approach for Explaining
Data
Amnon Catav, Boyang Fu, Yazeed Zoabi, Ahuva Libi Weiss Meilik, Noam Shomron, Jas
on Ernst, Sriram Sankararaman, Ran Gilad-Bachrach
In recent years, methods were proposed for assigning feature importance scores t
o measure the contribution of individual features. While in some cases the goal
is to understand a specific model, in many cases the goal is to understand the c
ontribution of certain properties (features) to a real-world phenomenon. Thus, a
 distinction has been made between feature importance scores that explain a mode
l and scores that explain the data. When explaining the data, machine learning m
odels are used as proxies in settings where conducting many real-world experimen
ts is expensive or prohibited. While existing feature importance scores show gre
at success in explaining models, we demonstrate their limitations when explainin
g the data, especially in the presence of correlations between features. Therefo
re, we develop a set of axioms to capture properties expected from a feature imp
ortance score when explaining data and prove that there exists only one score th
at satisfies all of them, the Marginal Contribution Feature Importance (MCI). We
 analyze the theoretical properties of this score function and demonstrate its m
erits empirically.
****
Disentangling syntax and semantics in the brain with deep networks
Charlotte Caucheteux, Alexandre Gramfort, Jean-Remi King
The activations of language transformers like GPT-2 have been shown to linearly
map onto brain activity during speech comprehension. However, the nature of thes
e activations remains largely unknown and presumably conflate distinct linguisti
c classes. Here, we propose a taxonomy to factorize the high-dimensional activat
ions of language models into four combinatorial classes: lexical, compositional,
 syntactic, and semantic representations. We then introduce a statistical method
 to decompose, through the lens of GPT-2’s activations, the brain activity of 34
5 subjects recorded with functional magnetic resonance imaging (fMRI) during the
 listening of 4.6 hours of narrated text. The results highlight two findings. F
irst, compositional representations recruit a more widespread cortical network t
han lexical ones, and encompass the bilateral temporal, parietal and prefrontal
cortices. Second, contrary to previous claims, syntax and semantics are not asso
ciated with separated modules, but, instead, appear to share a common and distri
buted neural substrate. Overall, this study introduces a versatile framework to
isolate, in the brain activity, the distributed representations of linguistic co

nstructs.
****
Fair Classification with Noisy Protected Attributes: A Framework with Provable Guarantees
L. Elisa Celis, Lingxiao Huang, Vijay Keswani, Nisheeth K. Vishnoi
We present an optimization framework for learning a fair classifier in the presence of noisy perturbations in the protected attributes. Compared to prior work, our framework can be employed with a very general class of linear and linear-fractional fairness constraints, can handle multiple, non-binary protected attributes, and outputs a classifier that comes with provable guarantees on both accuracy and fairness. Empirically, we show that our framework can be used to attain either statistical rate or false positive rate fairness guarantees with a minimal loss in accuracy, even when the noise is large, in two real-world datasets.
****
Best Model Identification: A Rested Bandit Formulation
Leonardo Cella, Massimiliano Pontil, Claudio Gentile
We introduce and analyze a best arm identification problem in the rested bandit setting, wherein arms are themselves learning algorithms whose expected losses decrease with the number of times the arm has been played. The shape of the expected loss functions is similar across arms, and is assumed to be available up to unknown parameters that have to be learned on the fly. We define a novel notion of regret for this problem, where we compare to the policy that always plays the arm having the smallest expected loss at the end of the game. We analyze an arm elimination algorithm whose regret vanishes as the time horizon increases. The actual rate of convergence depends in a detailed way on the postulated functional form of the expected losses. We complement our analysis with lower bounds, indicating strengths and limitations of the proposed solution.
****
Revisiting Rainbow: Promoting more insightful and inclusive deep reinforcement learning research
Johan Samir Obando Ceron, Pablo Samuel Castro
Since the introduction of DQN, a vast majority of reinforcement learning research has focused on reinforcement learning with deep neural networks as function approximators. New methods are typically evaluated on a set of environments that have now become standard, such as Atari 2600 games. While these benchmarks help standardize evaluation, their computational cost has the unfortunate side effect of widening the gap between those with ample access to computational resources, and those without. In this work we argue that, despite the community’s emphasis on large-scale environments, the traditional small-scale environments can still yield valuable scientific insights and can help reduce the barriers to entry for underprivileged communities. To substantiate our claims, we empirically revisit the paper which introduced the Rainbow algorithm [Hessel et al., 2018] and present some new insights into the algorithms used by Rainbow.
****
Learning Routines for Effective Off-Policy Reinforcement Learning
Edoardo Cetin, Oya Celiktutan
The performance of reinforcement learning depends upon designing an appropriate action space, where the effect of each action is measurable, yet, granular enough to permit flexible behavior. So far, this process involved non-trivial user choices in terms of the available actions and their execution frequency. We propose a novel framework for reinforcement learning that effectively lifts such constraints. Within our framework, agents learn effective behavior over a routine space: a new, higher-level action space, where each routine represents a set of ’equivalent’ sequences of granular actions with arbitrary length. Our routine space is learned end-to-end to facilitate the accomplishment of underlying off-policy reinforcement learning objectives. We apply our framework to two state-of-the-art off-policy algorithms and show that the resulting agents obtain relevant performance improvements while requiring fewer interactions with the environment per episode, improving computational efficiency.
****

Learning Node Representations Using Stationary Flow Prediction on Large Payment
and Cash Transaction Networks
Ciwan Ceylan, Salla Franzén, Florian T. Pokorny
Banks are required to analyse large transaction datasets as a part of the fight
against financial crime. Today, this analysis is either performed manually by do
main experts or using expensive feature engineering. Gradient flow analysis allo
ws for basic representation learning as node potentials can be inferred directly
 from network transaction data. However, the gradient model has a fundamental li
mitation: it cannot represent all types of of network flows. Furthermore, standa
rd methods for learning the gradient flow are not appropriate for flow signals t
hat span multiple orders of magnitude and contain outliers, i.e. transaction dat
a. In this work, the gradient model is extended to a gated version and we prove
that it, unlike the gradient model, is a universal approximator for flows on gra
phs. To tackle the mentioned challenges of transaction data, we propose a multi-
scale and outlier robust loss function based on the Student-t log-likelihood. Et
hereum transaction data is used for evaluation and the gradient models outperfor
m MLP models using hand-engineered and node2vec features in terms of relative er
ror. These results extend to 60 synthetic datasets, with experiments also showin
g that the gated gradient model learns qualitative information about the underly
ing synthetic generative flow distributions.
****
GRAND: Graph Neural Diffusion
Ben Chamberlain, James Rowbottom, Maria I Gorinova, Michael Bronstein, Stefan We
bb, Emanuele Rossi
We present Graph Neural Diffusion (GRAND) that approaches deep learning on graph
s as a continuous diffusion process and treats Graph Neural Networks (GNNs) as d
iscretisations of an underlying PDE. In our model, the layer structure and topol
ogy correspond to the discretisation choices of temporal and spatial operators.
Our approach allows a principled development of a broad new class of GNNs that a
re able to address the common plights of graph learning models such as depth, ov
ersmoothing, and bottlenecks. Key to the success of our models are stability wit
h respect to perturbations in the data and this is addressed for both implicit a
nd explicit discretisation schemes. We develop linear and nonlinear versions of
GRAND, which achieve competitive results on many standard graph benchmarks.
****
HoroPCA: Hyperbolic Dimensionality Reduction via Horospherical Projections
Ines Chami, Albert Gu, Dat P Nguyen, Christopher Re
This paper studies Principal Component Analysis (PCA) for data lying in hyperbol
ic spaces. Given directions, PCA relies on: (1) a parameterization of subspaces
spanned by these directions, (2) a method of projection onto subspaces that pres
erves information in these directions, and (3) an objective to optimize, namely
the variance explained by projections. We generalize each of these concepts to t
he hyperbolic space and propose HoroPCA, a method for hyperbolic dimensionality
reduction. By focusing on the core problem of extracting principal directions, H
oroPCA theoretically better preserves information in the original data such as d
istances, compared to previous generalizations of PCA. Empirically, we validate
that HoroPCA outperforms existing dimensionality reduction methods, significantl
y reducing error in distance preservation. As a data whitening method, it improv
es downstream classification by up to 3.9% compared to methods that don’t use wh
itening. Finally, we show that HoroPCA can be used to visualize hyperbolic data
in two dimensions.
****
Goal-Conditioned Reinforcement Learning with Imagined Subgoals
Elliot Chane-Sane, Cordelia Schmid, Ivan Laptev
Goal-conditioned reinforcement learning endows an agent with a large variety of
skills, but it often struggles to solve tasks that require more temporally exten
ded reasoning. In this work, we propose to incorporate imagined subgoals into po
licy learning to facilitate learning of complex tasks. Imagined subgoals are pre
dicted by a separate high-level policy, which is trained simultaneously with the
 policy and its critic. This high-level policy predicts intermediate states half

way to the goal using the value function as a reachability metric. We don’t require the policy to reach these subgoals explicitly. Instead, we use them to define a prior policy, and incorporate this prior into a KL-constrained policy iteration scheme to speed up and regularize learning. Imagined subgoals are used during policy learning, but not during test time, where we only apply the learned policy. We evaluate our approach on complex robotic navigation and manipulation tasks and show that it outperforms existing methods by a large margin.
****

Locally Private k-Means in One Round

Alisa Chang, Badih Ghazi, Ravi Kumar, Pasin Manurangsi

We provide an approximation algorithm for k-means clustering in the \emph{one-round} (aka \emph{non-interactive}) local model of differential privacy (DP). Our algorithm achieves an approximation ratio arbitrarily close to the best \emph{non private} approximation algorithm, improving upon previously known algorithms that only guarantee large (constant) approximation ratios. Furthermore, ours is the first constant-factor approximation algorithm for k-means that requires only \emph{one} round of communication in the local DP model, positively resolving an open question of Stemmer (SODA 2020). Our algorithmic framework is quite flexible; we demonstrate this by showing that it also yields a similar near-optimal approximation algorithm in the (one-round) shuffle DP model.
****

Modularity in Reinforcement Learning via Algorithmic Independence in Credit Assignment

Michael Chang, Sid Kaushik, Sergey Levine, Tom Griffiths

Many transfer problems require re-using previously optimal decisions for solving new tasks, which suggests the need for learning algorithms that can modify the mechanisms for choosing certain actions independently of those for choosing others. However, there is currently no formalism nor theory for how to achieve this kind of modular credit assignment. To answer this question, we define modular credit assignment as a constraint on minimizing the algorithmic mutual information among feedback signals for different decisions. We introduce what we call the modularity criterion for testing whether a learning algorithm satisfies this constraint by performing causal analysis on the algorithm itself. We generalize the recently proposed societal decision-making framework as a more granular formalism than the Markov decision process to prove that for decision sequences that do not contain cycles, certain single-step temporal difference action-value methods meet this criterion while all policy-gradient methods do not. Empirical evidence suggests that such action-value methods are more sample efficient than policy-gradient methods on transfer problems that require only sparse changes to a sequence of previously optimal decisions.
****

Image-Level or Object-Level? A Tale of Two Resampling Strategies for Long-Tailed Detection

Nadine Chang, Zhiding Yu, Yu-Xiong Wang, Animashree Anandkumar, Sanja Fidler, Jose M Alvarez

Training on datasets with long-tailed distributions has been challenging for major recognition tasks such as classification and detection. To deal with this challenge, image resampling is typically introduced as a simple but effective approach. However, we observe that long-tailed detection differs from classification since multiple classes may be present in one image. As a result, image resampling alone is not enough to yield a sufficiently balanced distribution at the object-level. We address object-level resampling by introducing an object-centric sampling strategy based on a dynamic, episodic memory bank. Our proposed strategy has two benefits: 1) convenient object-level resampling without significant extra computation, and 2) implicit feature-level augmentation from model updates. We show that image-level and object-level resamplings are both important, and thus unify them with a joint resampling strategy. Our method achieves state-of-the-art performance on the rare categories of LVIS, with 1.89% and 3.13% relative improvements over Forest R-CNN on detection and instance segmentation.
****

DeepWalking Backwards: From Embeddings Back to Graphs
Sudhanshu Chanpuriya, Cameron Musco, Konstantinos Sotiropoulos, Charalampos Tsou
rakakis
Low-dimensional node embeddings play a key role in analyzing graph datasets. How
ever, little work studies exactly what information is encoded by popular embeddi
ng methods, and how this information correlates with performance in downstream l
earning tasks. We tackle this question by studying whether embeddings can be inv
erted to (approximately) recover the graph used to generate them. Focusing on a
variant of the popular DeepWalk method \cite{PerozziAl-RfouSkiena:2014, QiuDongM
a:2018}, we present algorithms for accurate embedding inversion – i.e., from the
 low-dimensional embedding of a graph G, we can find a graph $\tilde G$ with a
 very similar embedding. We perform numerous experiments on real-world networks,
 observing that significant information about G, such as specific edges and bu
lk properties like triangle density, is often lost in $\tilde G$. However, commu
nity structure is often preserved or even enhanced. Our findings are a step towa
rds a more rigorous understanding of exactly what information embeddings encode
about the input graph, and why this information is useful for learning tasks.
****
Differentiable Spatial Planning using Transformers
Devendra Singh Chaplot, Deepak Pathak, Jitendra Malik
We consider the problem of spatial path planning. In contrast to the classical s
olutions which optimize a new plan from scratch and assume access to the full ma
p with ground truth obstacle locations, we learn a planner from the data in a di
fferentiable manner that allows us to leverage statistical regularities from pas
t data. We propose Spatial Planning Transformers (SPT), which given an obstacle
map learns to generate actions by planning over long-range spatial dependencies,
 unlike prior data-driven planners that propagate information locally via convol
utional structure in an iterative manner. In the setting where the ground truth
map is not known to the agent, we leverage pre-trained SPTs in an end-to-end fra
mework that has the structure of mapper and planner built into it which allows s
eamless generalization to out-of-distribution maps and goals. SPTs outperform pr
ior state-of-the-art differentiable planners across all the setups for both mani
pulation and navigation tasks, leading to an absolute improvement of 7-19%.
****
Solving Challenging Dexterous Manipulation Tasks With Trajectory Optimisation an
d Reinforcement Learning
Henry J Charlesworth, Giovanni Montana
Training agents to autonomously control anthropomorphic robotic hands has the po
tential to lead to systems capable of performing a multitude of complex manipula
tion tasks in unstructured and uncertain environments. In this work, we first in
troduce a suite of challenging simulated manipulation tasks where current reinfo
rcement learning and trajectory optimisation techniques perform poorly. These in
clude environments where two simulated hands have to pass or throw objects betwe
en each other, as well as an environment where the agent must learn to spin a lo
ng pen between its fingers. We then introduce a simple trajectory optimisation a
lgorithm that performs significantly better than existing methods on these envir
onments. Finally, on the most challenging “PenSpin" task, we combine sub-optimal
 demonstrations generated through trajectory optimisation with off-policy reinfo
rcement learning, obtaining performance that far exceeds either of these approac
hes individually. Videos of all of our results are available at: https://dextero
us-manipulation.github.io
****
Classification with Rejection Based on Cost-sensitive Classification
Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, Masashi Sugiyama
The goal of classification with rejection is to avoid risky misclassification in
 error-critical applications such as medical diagnosis and product inspection. I
n this paper, based on the relationship between classification with rejection an
d cost-sensitive classification, we propose a novel method of classification wit
h rejection by learning an ensemble of cost-sensitive classifiers, which satisfi
es all the following properties: (i) it can avoid estimating class-posterior pro

babilities, resulting in improved classification accuracy. (ii) it allows a flexible choice of losses including non-convex ones, (iii) it does not require complicated modifications when using different losses, (iv) it is applicable to both binary and multiclass cases, and (v) it is theoretically justifiable for any classification-calibrated loss. Experimental results demonstrate the usefulness of our proposed approach in clean-labeled, noisy-labeled, and positive-unlabeled classification.
****
Actionable Models: Unsupervised Offline Reinforcement Learning of Robotic Skills
Yevgen Chebotar, Karol Hausman, Yao Lu, Ted Xiao, Dmitry Kalashnikov, Jacob Varley, Alex Irpan, Benjamin Eysenbach, Ryan C Julian, Chelsea Finn, Sergey Levine
We consider the problem of learning useful robotic skills from previously collected offline data without access to manually specified rewards or additional online exploration, a setting that is becoming increasingly important for scaling robot learning by reusing past robotic data. In particular, we propose the objective of learning a functional understanding of the environment by learning to reach any goal state in a given dataset. We employ goal-conditioned Q-learning with hindsight relabeling and develop several techniques that enable training in a particularly challenging offline setting. We find that our method can operate on high-dimensional camera images and learn a variety of skills on real robots that generalize to previously unseen scenes and objects. We also show that our method can learn to reach long-horizon goals across multiple episodes through goal chaining, and learn rich representations that can help with downstream tasks through pre-training or auxiliary objectives.
****
Unified Robust Semi-Supervised Variational Autoencoder
Xu Chen
In this paper, we propose a novel noise-robust semi-supervised deep generative model by jointly tackling noisy labels and outliers simultaneously in a unified robust semi-supervised variational autoencoder (URSVAE). Typically, the uncertainty of of input data is characterized by placing uncertainty prior on the parameters of the probability density distributions in order to ensure the robustness of the variational encoder towards outliers. Subsequently, a noise transition model is integrated naturally into our model to alleviate the detrimental effects of noisy labels. Moreover, a robust divergence measure is employed to further enhance the robustness, where a novel variational lower bound is derived and optimized to infer the network parameters. By proving the influence function on the proposed evidence lower bound is bounded, the enormous potential of the proposed model in the classification in the presence of the compound noise is demonstrated. The experimental results highlight the superiority of the proposed framework by the evaluating on image classification tasks and comparing with the state-of-the-art approaches.
****
Unsupervised Learning of Visual 3D Keypoints for Control
Boyuan Chen, Pieter Abbeel, Deepak Pathak
Learning sensorimotor control policies from high-dimensional images crucially relies on the quality of the underlying visual representations. Prior works show that structured latent space such as visual keypoints often outperforms unstructured representations for robotic control. However, most of these representations, whether structured or unstructured are learned in a 2D space even though the control tasks are usually performed in a 3D environment. In this work, we propose a framework to learn such a 3D geometric structure directly from images in an end-to-end unsupervised manner. The input images are embedded into latent 3D keypoints via a differentiable encoder which is trained to optimize both a multi-view consistency loss and downstream task objective. These discovered 3D keypoints tend to meaningfully capture robot joints as well as object movements in a consistent manner across both time and 3D space. The proposed approach outperforms prior state-of-art methods across a variety of reinforcement learning benchmarks. Code and videos at https://buoyancy99.github.io/unsup-3d-keypoints/.
****

Integer Programming for Causal Structure Learning in the Presence of Latent Variables
Rui Chen, Sanjeeb Dash, Tian Gao
The problem of finding an ancestral acyclic directed mixed graph (ADMG) that represents the causal relationships between a set of variables is an important area of research on causal inference. Most existing score-based structure learning methods focus on learning directed acyclic graph (DAG) models without latent variables. A number of score-based methods have recently been proposed for the ADMG learning, yet they are heuristic in nature and do not guarantee an optimal solution. We propose a novel exact score-based method that solves an integer programming (IP) formulation and returns a score-maximizing ancestral ADMG for a set of continuous variables that follow a multivariate Gaussian distribution. We generalize the state-of-the-art IP model for DAG learning problems and derive new classes of valid inequalities to formulate an IP model for ADMG learning. Empirically, our model can be solved efficiently for medium-sized problems and achieves better accuracy than state-of-the-art score-based methods as well as benchmark constraint-based methods.
****
Improved Corruption Robust Algorithms for Episodic Reinforcement Learning
Yifang Chen, Simon Du, Kevin Jamieson
We study episodic reinforcement learning under unknown adversarial corruptions in both the rewards and the transition probabilities of the underlying system. We propose new algorithms which, compared to the existing results in \cite{lykouris2020corruption}, achieve strictly better regret bounds in terms of total corruptions for the tabular setting. To be specific, firstly, our regret bounds depend on more precise numerical values of total rewards corruptions and transition corruptions, instead of only on the total number of corrupted episodes. Secondly, our regret bounds are the first of their kind in the reinforcement learning setting to have the number of corruptions show up additively with respect to $\min\{ \sqrt{T},\text{PolicyGapComplexity} \}$ rather than multiplicatively. Our results follow from a general algorithmic framework that combines corruption-robust policy elimination meta-algorithms, and plug-in reward-free exploration sub-algorithms. Replacing the meta-algorithm or sub-algorithm may extend the framework to address other corrupted settings with potentially more structure.
****
Scalable Computations of Wasserstein Barycenter via Input Convex Neural Networks
Jiaojiao Fan, Amirhossein Taghvaei, Yongxin Chen
Wasserstein Barycenter is a principled approach to represent the weighted mean of a given set of probability distributions, utilizing the geometry induced by optimal transport. In this work, we present a novel scalable algorithm to approximate the Wasserstein Barycenters aiming at high-dimensional applications in machine learning. Our proposed algorithm is based on the Kantorovich dual formulation of the Wasserstein-2 distance as well as a recent neural network architecture, input convex neural network, that is known to parametrize convex functions. The distinguishing features of our method are: i) it only requires samples from the marginal distributions; ii) unlike the existing approaches, it represents the Barycenter with a generative model and can thus generate infinite samples from the barycenter without querying the marginal distributions; iii) it works similar to Generative Adversarial Model in one marginal case. We demonstratethe efficacy of our algorithm by comparing it with the state-of-art methods in multiple experiments.
****
Neural Feature Matching in Implicit 3D Representations
Yunlu Chen, Basura Fernando, Hakan Bilen, Thomas Mensink, Efstratios Gavves
Recently, neural implicit functions have achieved impressive results for encoding 3D shapes. Conditioning on low-dimensional latent codes generalises a single implicit function to learn shared representation space for a variety of shapes, with the advantage of smooth interpolation. While the benefits from the global latent space do not correspond to explicit points at local level, we propose to track the continuous point trajectory by matching implicit features with the laten

t code interpolating between shapes, from which we corroborate the hierarchical functionality of the deep implicit functions, where early layers map the latent code to fitting the coarse shape structure, and deeper layers further refine the shape details. Furthermore, the structured representation space of implicit functions enables to apply feature matching for shape deformation, with the benefits to handle topology and semantics inconsistency, such as from an armchair to a chair with no arms, without explicit flow functions or manual annotations.
****
Decentralized Riemannian Gradient Descent on the Stiefel Manifold
Shixiang Chen, Alfredo Garcia, Mingyi Hong, Shahin Shahrampour
We consider a distributed non-convex optimization where a network of agents aims at minimizing a global function over the Stiefel manifold. The global function is represented as a finite sum of smooth local functions, where each local function is associated with one agent and agents communicate with each other over an undirected connected graph. The problem is non-convex as local functions are possibly non-convex (but smooth) and the Steifel manifold is a non-convex set. We present a decentralized Riemannian stochastic gradient method (DRSGD) with the convergence rate of $\mathcal{O}(1/\sqrt{K})$ to a stationary point. To have exact convergence with constant stepsize, we also propose a decentralized Riemannian gradient tracking algorithm (DRGTA) with the convergence rate of $\mathcal{O}(1/K)$ to a stationary point. We use multi-step consensus to preserve the iteration in the local (consensus) region. DRGTA is the first decentralized algorithm with exact convergence for distributed optimization on Stiefel manifold.
****
Learning Self-Modulating Attention in Continuous Time Space with Applications to Sequential Recommendation
Chao Chen, Haoyu Geng, Nianzu Yang, Junchi Yan, Daiyue Xue, Jianping Yu, Xiaokang Yang
User interests are usually dynamic in the real world, which poses both theoretical and practical challenges for learning accurate preferences from rich behavior data. Among existing user behavior modeling solutions, attention networks are widely adopted for its effectiveness and relative simplicity. Despite being extensively studied, existing attentions still suffer from two limitations: i) conventional attentions mainly take into account the spatial correlation between user behaviors, regardless the distance between those behaviors in the continuous time space; and ii) these attentions mostly provide a dense and undistinguished distribution over all past behaviors then attentively encode them into the output latent representations. This is however not suitable in practical scenarios where a user’s future actions are relevant to a small subset of her/his historical behaviors. In this paper, we propose a novel attention network, named \textit{self-modulating attention}, that models the complex and non-linearly evolving dynamic user preferences. We empirically demonstrate the effectiveness of our method on top-N sequential recommendation tasks, and the results on three large-scale real-world datasets show that our model can achieve state-of-the-art performance.
****
Mandoline: Model Evaluation under Distribution Shift
Mayee Chen, Karan Goel, Nimit S Sohoni, Fait Poms, Kayvon Fatahalian, Christopher Re
Machine learning models are often deployed in different settings than they were trained and validated on, posing a challenge to practitioners who wish to predict how well the deployed model will perform on a target distribution. If an unlabeled sample from the target distribution is available, along with a labeled sample from a possibly different source distribution, standard approaches such as importance weighting can be applied to estimate performance on the target. However, importance weighting struggles when the source and target distributions have non-overlapping support or are high-dimensional. Taking inspiration from fields such as epidemiology and polling, we develop Mandoline, a new evaluation framework that mitigates these issues. Our key insight is that practitioners may have prior knowledge about the ways in which the distribution shifts, which we can use to better guide the importance weighting procedure. Specifically, users write si

mple "slicing functions" {–} noisy, potentially correlated binary functions intended to capture possible axes of distribution shift {–} to compute reweighted performance estimates. We further describe a density ratio estimation framework for the slices and show how its estimation error scales with slice quality and dataset size. Empirical validation on NLP and vision tasks shows that Mandoline can estimate performance on the target distribution up to 3x more accurately compared to standard baselines.
****
Order Matters: Probabilistic Modeling of Node Sequence for Graph Generation
Xiaohui Chen, Xu Han, Jiajing Hu, Francisco Ruiz, Liping Liu
A graph generative model defines a distribution over graphs. Typically, the model consists of a sequential process that creates and adds nodes and edges. Such sequential process defines an ordering of the nodes in the graph. The computation of the model’s likelihood requires to marginalize the node orderings; this makes maximum likelihood estimation (MLE) challenging due to the (factorial) number of possible permutations. In this work, we provide an expression for the likelihood of a graph generative model and show that its calculation is closely related to the problem of graph automorphism. In addition, we derive a variational inference (VI) algorithm for fitting a graph generative model that is based on the maximization of a variational bound of the log-likelihood. This allows the model to be trained with node orderings from the approximate posterior instead of ad-hoc orderings. Our experiments show that our log-likelihood bound is significantly tighter than the bound of previous schemes. The models fitted with the VI algorithm are able to generate high-quality graphs that match the structures of target graphs not seen during training.
****
CARTL: Cooperative Adversarially-Robust Transfer Learning
Dian Chen, Hongxin Hu, Qian Wang, Li Yinli, Cong Wang, Chao Shen, Qi Li
Transfer learning eases the burden of training a well-performed model from scratch, especially when training data is scarce and computation power is limited. In deep learning, a typical strategy for transfer learning is to freeze the early layers of a pre-trained model and fine-tune the rest of its layers on the target domain. Previous work focuses on the accuracy of the transferred model but neglects the transfer of adversarial robustness. In this work, we first show that transfer learning improves the accuracy on the target domain but degrades the inherited robustness of the target model. To address such a problem, we propose a novel cooperative adversarially-robust transfer learning (CARTL) by pre-training the model via feature distance minimization and fine-tuning the pre-trained model with non-expansive fine-tuning for target domain tasks. Empirical results show that CARTL improves the inherited robustness by about 28% at most compared with the baseline with the same degree of accuracy. Furthermore, we study the relationship between the batch normalization (BN) layers and the robustness in the context of transfer learning, and we reveal that freezing BN layers can further boost the robustness transfer.
****
Finding the Stochastic Shortest Path with Low Regret: the Adversarial Cost and Unknown Transition Case
Liyu Chen, Haipeng Luo
We make significant progress toward the stochastic shortest path problem with adversarial costs and unknown transition. Specifically, we develop algorithms that achieve $O(\sqrt{S^2ADT_\star K})$ regret for the full-information setting and $O(\sqrt{S^3A^2DT_\star K})$ regret for the bandit feedback setting, where D is the diameter, $T_\star$ is the expected hitting time of the optimal policy, S is the number of states, A is the number of actions, and K is the number of episodes. Our work strictly improves (Rosenberg and Mansour, 2020) in the full information setting, extends (Chen et al., 2020) from known transition to unknown transition, and is also the first to consider the most challenging combination: bandit feedback with adversarial costs and unknown transition. To remedy the gap between our upper bounds and the current best lower bounds constructed via a stochastically oblivious adversary, we also propose algorithms with near-optima

l regret for this special case.
****
SpreadsheetCoder: Formula Prediction from Semi-structured Context
Xinyun Chen, Petros Maniatis, Rishabh Singh, Charles Sutton, Hanjun Dai, Max Lin, Denny Zhou
Spreadsheet formula prediction has been an important program synthesis problem with many real-world applications. Previous works typically utilize input-output examples as the specification for spreadsheet formula synthesis, where each input-output pair simulates a separate row in the spreadsheet. However, this formulation does not fully capture the rich context in real-world spreadsheets. First, spreadsheet data entries are organized as tables, thus rows and columns are not necessarily independent from each other. In addition, many spreadsheet tables include headers, which provide high-level descriptions of the cell data. However, previous synthesis approaches do not consider headers as part of the specification. In this work, we present the first approach for synthesizing spreadsheet formulas from tabular context, which includes both headers and semi-structured tabular data. In particular, we propose SpreadsheetCoder, a BERT-based model architecture to represent the tabular context in both row-based and column-based formats. We train our model on a large dataset of spreadsheets, and demonstrate that SpreadsheetCoder achieves top-1 prediction accuracy of 42.51%, which is a considerable improvement over baselines that do not employ rich tabular context. Compared to the rule-based system, SpreadsheetCoder assists 82% more users in composing formulas on Google Sheets.
****
Large-Margin Contrastive Learning with Distance Polarization Regularizer
Shuo Chen, Gang Niu, Chen Gong, Jun Li, Jian Yang, Masashi Sugiyama
\emph{Contrastive learning} (CL) pretrains models in a pairwise manner, where given a data point, other data points are all regarded as dissimilar, including some that are \emph{semantically} similar. The issue has been addressed by properly weighting similar and dissimilar pairs as in \emph{positive-unlabeled learning}, so that the objective of CL is \emph{unbiased} and CL is \emph{consistent}. However, in this paper, we argue that this great solution is still not enough: its weighted objective \emph{hides} the issue where the semantically similar pairs are still pushed away; as CL is pretraining, this phenomenon is not our desideratum and might affect downstream tasks. To this end, we propose \emph{large-margin contrastive learning} (LMCL) with \emph{distance polarization regularizer}, motivated by the distribution characteristic of pairwise distances in \emph{metric learning}. In LMCL, we can distinguish between \emph{intra-cluster} and \emph{inter-cluster} pairs, and then only push away inter-cluster pairs, which \emph{solves} the above issue explicitly. Theoretically, we prove a tighter error bound for LMCL; empirically, the superiority of LMCL is demonstrated across multiple domains, \emph{i.e.}, image classification, sentence representation, and reinforcement learning.
****
Z-GCNETs: Time Zigzags at Graph Convolutional Networks for Time Series Forecasting
Yuzhou Chen, Ignacio Segovia, Yulia R. Gel
There recently has been a surge of interest in developing a new class of deep learning (DL) architectures that integrate an explicit time dimension as a fundamental building block of learning and representation mechanisms. In turn, many recent results show that topological descriptors of the observed data, encoding information on the shape of the dataset in a topological space at different scales, that is, persistent homology of the data, may contain important complementary information, improving both performance and robustness of DL. As convergence of these two emerging ideas, we propose to enhance DL architectures with the most salient time-conditioned topological information of the data and introduce the concept of zigzag persistence into time-aware graph convolutional networks (GCNs). Zigzag persistence provides a systematic and mathematically rigorous framework to track the most important topological features of the observed data that tend to manifest themselves over time. To integrate the extracted time-conditioned top

ological descriptors into DL, we develop a new topological summary, zigzag persistence image, and derive its theoretical stability guarantees. We validate the new GCNs with a time-aware zigzag topological layer (Z-GCNETs), in application to traffic forecasting and Ethereum blockchain price prediction. Our results indicate that Z-GCNET outperforms 13 state-of-the-art methods on 4 time series datasets.
****
A Unified Lottery Ticket Hypothesis for Graph Neural Networks
Tianlong Chen, Yongduo Sui, Xuxi Chen, Aston Zhang, Zhangyang Wang
With graphs rapidly growing in size and deeper graph neural networks (GNNs) emerging, the training and inference of GNNs become increasingly expensive. Existing network weight pruning algorithms cannot address the main space and computational bottleneck in GNNs, caused by the size and connectivity of the graph. To this end, this paper first presents a unified GNN sparsification (UGS) framework that simultaneously prunes the graph adjacency matrix and the model weights, for effectively accelerating GNN inference on large-scale graphs. Leveraging this new tool, we further generalize the recently popular lottery ticket hypothesis to GNNs for the first time, by defining a graph lottery ticket (GLT) as a pair of core sub-dataset and sparse sub-network, which can be jointly identified from the original GNN and the full dense graph by iteratively applying UGS. Like its counterpart in convolutional neural networks, GLT can be trained in isolation to match the performance of training with the full model and graph, and can be drawn from both randomly initialized and self-supervised pre-trained GNNs. Our proposal has been experimentally verified across various GNN architectures and diverse tasks, on both small-scale graph datasets (Cora, Citeseer and PubMed), and large-scale datasets from the challenging Open Graph Benchmark (OGB). Specifically, for node classification, our found GLTs achieve the same accuracies with 20% 98% MACs saving on small graphs and 25% 85% MACs saving on large ones. For link prediction, GLTs lead to 48% 97% and 70% MACs saving on small and large graph datasets, respectively, without compromising predictive performance. Codes are at https://github.com/VITA-Group/Unified-LTH-GNN.
****
Network Inference and Influence Maximization from Samples
Wei Chen, Xiaoming Sun, Jialin Zhang, Zhijie Zhang
Influence maximization is the task of selecting a small number of seed nodes in a social network to maximize the spread of the influence from these seeds, and it has been widely investigated in the past two decades. In the canonical setting, the whole social network as well as its diffusion parameters is given as input. In this paper, we consider the more realistic sampling setting where the network is unknown and we only have a set of passively observed cascades that record the set of activated nodes at each diffusion step. We study the task of influence maximization from these cascade samples (IMS), and present constant approximation algorithms for this task under mild conditions on the seed set distribution. To achieve the optimization goal, we also provide a novel solution to the network inference problem, that is, learning diffusion parameters and the network structure from the cascade data. Comparing with prior solutions, our network inference algorithm requires weaker assumptions and does not rely on maximum-likelihood estimation and convex programming. Our IMS algorithms enhance the learning-and-then-optimization approach by allowing a constant approximation ratio even when the diffusion parameters are hard to learn, and we do not need any assumption related to the network structure or diffusion parameters.
****
Data-driven Prediction of General Hamiltonian Dynamics via Learning Exactly-Symplectic Maps
Renyi Chen, Molei Tao
We consider the learning and prediction of nonlinear time series generated by a latent symplectic map. A special case is (not necessarily separable) Hamiltonian systems, whose solution flows give such symplectic maps. For this special case, both generic approaches based on learning the vector field of the latent ODE and specialized approaches based on learning the Hamiltonian that generates the ve

ctor field exist. Our method, however, is different as it does not rely on the v
ector field nor assume its existence; instead, it directly learns the symplectic
 evolution map in discrete time. Moreover, we do so by representing the symplect
ic map via a generating function, which we approximate by a neural network (henc
e the name GFNN). This way, our approximation of the evolution map is always \em
ph{exactly} symplectic. This additional geometric structure allows the local pre
diction error at each step to accumulate in a controlled fashion, and we will pr
ove, under reasonable assumptions, that the global prediction error grows at mos
t \emph{linearly} with long prediction time, which significantly improves an oth
erwise exponential growth. In addition, as a map-based and thus purely data-driv
en method, GFNN avoids two additional sources of inaccuracies common in vector-f
ield based approaches, namely the error in approximating the vector field by fin
ite difference of the data, and the error in numerical integration of the vector
 field for making predictions. Numerical experiments further demonstrate our cla
ims.
****
Analysis of stochastic Lanczos quadrature for spectrum approximation
Tyler Chen, Thomas Trogdon, Shashanka Ubaru
The cumulative empirical spectral measure (CESM) $\Phi[\mathbf{A}] : \mathbb{R}
\to [0,1]$ of a $n\times n$ symmetric matrix $\mathbf{A}$ is defined as the frac
tion of eigenvalues of $\mathbf{A}$ less than a given threshold, i.e., $\Phi[\ma
thbf{A}](x) := \sum_{i=1}^{n} \frac{1}{n} {\large\unicode{x1D7D9}}[\lambda_i[\m
athbf{A}]\leq x]$. Spectral sums $\operatorname{tr}(f[\mathbf{A}])$ can be compu
ted as the Riemann–Stieltjes integral of f against $\Phi[\mathbf{A}]$, so the
task of estimating CESM arises frequently in a number of applications, including
 machine learning. We present an error analysis for stochastic Lanczos quadratur
e (SLQ). We show that SLQ obtains an approximation to the CESM within a Wasserst
ein distance of $t \: | \lambda_{\text{max}}[\mathbf{A}] - \lambda_{\text{min}}[
\mathbf{A}] |$ with probability at least $1-\eta$, by applying the Lanczos algor
ithm for $\lceil 12 t^{-1} + \frac{1}{2} \rceil$ iterations to $\lceil 4 (n+2)
^{-1}t^{-2} \ln(2n\eta^{-1}) \rceil$ vectors sampled independently and uniformly
 from the unit sphere. We additionally provide (matrix-dependent) a posteriori e
rror bounds for the Wasserstein and Kolmogorov–Smirnov distances between the out
put of this algorithm and the true CESM. The quality of our bounds is demonstrat
ed using numerical experiments.
****
Large-Scale Multi-Agent Deep FBSDEs
Tianrong Chen, Ziyi O Wang, Ioannis Exarchos, Evangelos Theodorou
In this paper we present a scalable deep learning framework for finding Markovia
n Nash Equilibria in multi-agent stochastic games using fictitious play. The mot
ivation is inspired by theoretical analysis of Forward Backward Stochastic Diffe
rential Equations and their implementation in a deep learning setting, which is
the source of our algorithm’s sample efficiency improvement. By taking advantage
 of the permutation-invariant property of agents in symmetric games, the scalabi
lity and performance is further enhanced significantly. We showcase superior per
formance of our framework over the state-of-the-art deep fictitious play algorit
hm on an inter-bank lending/borrowing problem in terms of multiple metrics. More
 importantly, our approach scales up to 3000 agents in simulation, a scale which
, to the best of our knowledge, represents a new state-of-the-art. We also demon
strate the applicability of our framework in robotics on a belief space autonomo
us racing problem.
****
Representation Subspace Distance for Domain Adaptation Regression
Xinyang Chen, Sinan Wang, Jianmin Wang, Mingsheng Long
Regression, as a counterpart to classification, is a major paradigm with a wide
range of applications. Domain adaptation regression extends it by generalizing a
 regressor from a labeled source domain to an unlabeled target domain. Existing
domain adaptation regression methods have achieved positive results limited only
 to the shallow regime. A question arises: Why learning invariant representation
s in the deep regime less pronounced? A key finding of this paper is that classi

fication is robust to feature scaling but regression is not, and aligning the distributions of deep representations will alter feature scale and impede domain adaptation regression. Based on this finding, we propose to close the domain gap through orthogonal bases of the representation spaces, which are free from feature scaling. Inspired by Riemannian geometry of Grassmann manifold, we define a geometrical distance over representation subspaces and learn deep transferable representations by minimizing it. To avoid breaking the geometrical properties of deep representations, we further introduce the bases mismatch penalization to match the ordering of orthogonal bases across representation subspaces. Our method is evaluated on three domain adaptation regression benchmarks, two of which are introduced in this paper. Our method outperforms the state-of-the-art methods significantly, forming early positive results in the deep regime.
****
Overcoming Catastrophic Forgetting by Bayesian Generative Regularization
Pei-Hung Chen, Wei Wei, Cho-Jui Hsieh, Bo Dai
In this paper, we propose a new method to over-come catastrophic forgetting by adding generative regularization to Bayesian inference frame-work. Bayesian method provides a general frame-work for continual learning. We could further construct a generative regularization term for all given classification models by leveraging energy-based models and Langevin dynamic sampling to enrich the features learned in each task. By combining discriminative and generative loss together, we empirically show that the proposed method outperforms state-of-the-art methods on a variety of tasks, avoiding catastrophic forgetting in continual learning. In particular, the proposed method outperforms baseline methods over 15%on the Fashion-MNIST dataset and 10%on the CUB dataset.
****
Cyclically Equivariant Neural Decoders for Cyclic Codes
Xiangyu Chen, Min Ye
Neural decoders were introduced as a generalization of the classic Belief Propagation (BP) decoding algorithms, where the Trellis graph in the BP algorithm is viewed as a neural network, and the weights in the Trellis graph are optimized by training the neural network. In this work, we propose a novel neural decoder for cyclic codes by exploiting their cyclically invariant property. More precisely, we impose a shift invariant structure on the weights of our neural decoder so that any cyclic shift of inputs results in the same cyclic shift of outputs. Extensive simulations with BCH codes and punctured Reed-Muller (RM) codes show that our new decoder consistently outperforms previous neural decoders when decoding cyclic codes. Finally, we propose a list decoding procedure that can significantly reduce the decoding error probability for BCH codes and punctured RM codes. For certain high-rate codes, the gap between our list decoder and the Maximum Likelihood decoder is less than 0.1dB. Code available at github.com/cyclicallyneuraldecoder
****
A Receptor Skeleton for Capsule Neural Networks
Jintai Chen, Hongyun Yu, Chengde Qian, Danny Z Chen, Jian Wu
In previous Capsule Neural Networks (CapsNets), routing algorithms often performed clustering processes to assemble the child capsules’ representations into parent capsules. Such routing algorithms were typically implemented with iterative processes and incurred high computing complexity. This paper presents a new capsule structure, which contains a set of optimizable receptors and a transmitter is devised on the capsule’s representation. Specifically, child capsules’ representations are sent to the parent capsules whose receptors match well the transmitters of the child capsules’ representations, avoiding applying computationally complex routing algorithms. To ensure the receptors in a CapsNet work cooperatively, we build a skeleton to organize the receptors in different capsule layers in a CapsNet. The receptor skeleton assigns a share-out objective for each receptor, making the CapsNet perform as a hierarchical agglomerative clustering process. Comprehensive experiments verify that our approach facilitates efficient clustering processes, and CapsNets with our approach significantly outperform CapsNets with previous routing algorithms on image classification, affine transformatio

n generalization, overlapped object recognition, and representation semantic decoupling.
****
Accelerating Gossip SGD with Periodic Global Averaging
Yiming Chen, Kun Yuan, Yingya Zhang, Pan Pan, Yinghui Xu, Wotao Yin
Communication overhead hinders the scalability of large-scale distributed training. Gossip SGD, where each node averages only with its neighbors, is more communication-efficient than the prevalent parallel SGD. However, its convergence rate is reversely proportional to quantity $1-\beta$ which measures the network connectivity. On large and sparse networks where $1-\beta \to 0$, Gossip SGD requires more iterations to converge, which offsets against its communication benefit. This paper introduces Gossip-PGA, which adds Periodic Global Averaging to accelerate Gossip SGD. Its transient stage, i.e., the iterations required to reach asymptotic linear speedup stage, improves from $\Omega(\beta^4 n^3/(1-\beta)^4)$ to $\Omega(\beta^4 n^3 H^4)$ for non-convex problems. The influence of network topology in Gossip-PGA can be controlled by the averaging period H. Its transient-stage complexity is also superior to local SGD which has order $\Omega(n^3 H^4)$. Empirical results of large-scale training on image classification (ResNet50) and language modeling (BERT) validate our theoretical findings.
****
ActNN: Reducing Training Memory Footprint via 2-Bit Activation Compressed Training
Jianfei Chen, Lianmin Zheng, Zhewei Yao, Dequan Wang, Ion Stoica, Michael Mahoney, Joseph Gonzalez
The increasing size of neural network models has been critical for improvements in their accuracy, but device memory is not growing at the same rate. This creates fundamental challenges for training neural networks within limited memory environments. In this work, we propose ActNN, a memory-efficient training framework that stores randomly quantized activations for back propagation. We prove the convergence of ActNN for general network architectures, and we characterize the impact of quantization on the convergence via an exact expression for the gradient variance. Using our theory, we propose novel mixed-precision quantization strategies that exploit the activation’s heterogeneity across feature dimensions, samples, and layers. These techniques can be readily applied to existing dynamic graph frameworks, such as PyTorch, simply by substituting the layers. We evaluate ActNN on mainstream computer vision models for classification, detection, and segmentation tasks. On all these tasks, ActNN compresses the activation to 2 bits on average, with negligible accuracy loss. ActNN reduces the memory footprint of the activation by 12x, and it enables training with a 6.6x to 14x larger batch size.
****
SPADE: A Spectral Method for Black-Box Adversarial Robustness Evaluation
Wuxinlin Cheng, Chenhui Deng, Zhiqiang Zhao, Yaohui Cai, Zhiru Zhang, Zhuo Feng
A black-box spectral method is introduced for evaluating the adversarial robustness of a given machine learning (ML) model. Our approach, named SPADE, exploits bijective distance mapping between the input/output graphs constructed for approximating the manifolds corresponding to the input/output data. By leveraging the generalized Courant-Fischer theorem, we propose a SPADE score for evaluating the adversarial robustness of a given model, which is proved to be an upper bound of the best Lipschitz constant under the manifold setting. To reveal the most non-robust data samples highly vulnerable to adversarial attacks, we develop a spectral graph embedding procedure leveraging dominant generalized eigenvectors. This embedding step allows assigning each data point a robustness score that can be further harnessed for more effective adversarial training of ML models. Our experiments show promising empirical results for neural networks trained with the MNIST and CIFAR-10 data sets.
****
Self-supervised and Supervised Joint Training for Resource-rich Machine Translation
Yong Cheng, Wei Wang, Lu Jiang, Wolfgang Macherey

Self-supervised pre-training of text representations has been successfully appli
ed to low-resource Neural Machine Translation (NMT). However, it usually fails t
o achieve notable gains on resource-rich NMT. In this paper, we propose a joint
training approach, F2-XEnDec, to combine self-supervised and supervised learning
to optimize NMT models. To exploit complementary self-supervised signals for su
pervised learning, NMT models are trained on examples that are interbred from mo
nolingual and parallel sentences through a new process called crossover encoder-
decoder. Experiments on two resource-rich translation benchmarks, WMT’14 English
-German and WMT’14 English-French, demonstrate that our approach achieves substa
ntial improvements over several strong baseline methods and obtains a new state
of the art of 46.19 BLEU on English-French when incorporating back translation.
Results also show that our approach is capable of improving model robustness to
input perturbations such as code-switching noise which frequently appears on the
social media.
****
Exact Optimization of Conformal Predictors via Incremental and Decremental Learn
ing
Giovanni Cherubin, Konstantinos Chatzikokolakis, Martin Jaggi
Conformal Predictors (CP) are wrappers around ML models, providing error guarant
ees under weak assumptions on the data distribution. They are suitable for a wid
e range of problems, from classification and regression to anomaly detection. Un
fortunately, their very high computational complexity limits their applicability
to large datasets. In this work, we show that it is possible to speed up a CP c
lassifier considerably, by studying it in conjunction with the underlying ML met
hod, and by exploiting incremental&decremental learning. For methods such as k-N
N, KDE, and kernel LS-SVM, our approach reduces the running time by one order of
magnitude, whilst producing exact solutions. With similar ideas, we also achiev
e a linear speed up for the harder case of bootstrapping. Finally, we extend the
se techniques to improve upon an optimization of k-NN CP for regression. We eval
uate our findings empirically, and discuss when methods are suitable for CP opti
mization.
****
Problem Dependent View on Structured Thresholding Bandit Problems
James Cheshire, Pierre Menard, Alexandra Carpentier
We investigate the \textit{problem dependent regime} in the stochastic \emph{Thr
esholding Bandit problem} (\tbp) under several \emph{shape constraints}. In the
\tbp the objective of the learner is to output, after interacting with the envir
onment, the set of arms whose means are above a given threshold. The vanilla, un
structured, case is already well studied in the literature. Taking K as the nu
mber of arms, we consider the case where (i) the sequence of arm’s means $(\mu_k
){k=1}^K$ is monotonically increasing (\textit{MTBP}) and (ii) the case where $(
\mu_k){k=1}^K$ is concave (\textit{CTBP}). We consider both cases in the \emph{p
roblem dependent} regime and study the probability of error - i.e. the probabili
ty to mis-classify at least one arm. In the fixed budget setting, we provide nea
rly matching upper and lower bounds for the probability of error in both the con
cave and monotone settings, as well as associated algorithms. Of interest, is th
at for both the monotone and concave cases, optimal bounds on probability of err
or are of the same order as those for the two armed bandit problem.
****
Online Optimization in Games via Control Theory: Connecting Regret, Passivity an
d Poincaré Recurrence
Yun Kuen Cheung, Georgios Piliouras
We present a novel control-theoretic understanding of online optimization and le
arning in games, via the notion of passivity. Passivity is a fundamental concept
in control theory, which abstracts energy conservation and dissipation in physi
cal systems. It has become a standard tool in analysis of general feedback syste
ms, to which game dynamics belong. Our starting point is to show that all contin
uous-time Follow-the-Regularized-Leader (FTRL) dynamics, which include the well-
known Replicator Dynamic, are lossless, i.e. it is passive with no energy dissip
ation. Interestingly, we prove that passivity implies bounded regret, connecting

two fundamental primitives of control theory and online optimization. The observation of energy conservation in FTRL inspires us to present a family of lossless learning dynamics, each of which has an underlying energy function with a simple gradient structure. This family is closed under convex combination; as an immediate corollary, any convex combination of FTRL dynamics is lossless and thus has bounded regret. This allows us to extend the framework of Fox & Shamma [Games 2013] to prove not just global asymptotic stability results for game dynamics, but Poincar{é} recurrence results as well. Intuitively, when a lossless game (e.g. graphical constant-sum game) is coupled with lossless learning dynamic, their interconnection is also lossless, which results in a pendulum-like energy-preserving recurrent behavior, generalizing Piliouras & Shamma [SODA 2014] and Mertikopoulos et al. [SODA 2018].
****
Understanding and Mitigating Accuracy Disparity in Regression
Jianfeng Chi, Yuan Tian, Geoffrey J. Gordon, Han Zhao
With the widespread deployment of large-scale prediction systems in high-stakes domains, e.g., face recognition, criminal justice, etc., disparity on prediction accuracy between different demographic subgroups has called for fundamental understanding on the source of such disparity and algorithmic intervention to mitigate it. In this paper, we study the accuracy disparity problem in regression. To begin with, we first propose an error decomposition theorem, which decomposes the accuracy disparity into the distance between marginal label distributions and the distance between conditional representations, to help explain why such accuracy disparity appears in practice. Motivated by this error decomposition and the general idea of distribution alignment with statistical distances, we then propose an algorithm to reduce this disparity, and analyze its game-theoretic optima of the proposed objective functions. To corroborate our theoretical findings, we also conduct experiments on five benchmark datasets. The experimental results suggest that our proposed algorithms can effectively mitigate accuracy disparity while maintaining the predictive power of the regression models.
****
Private Alternating Least Squares: Practical Private Matrix Completion with Tighter Rates
Steve Chien, Prateek Jain, Walid Krichene, Steffen Rendle, Shuang Song, Abhradeep Thakurta, Li Zhang
We study the problem of differentially private (DP) matrix completion under user-level privacy. We design a joint differentially private variant of the popular Alternating-Least-Squares (ALS) method that achieves: i) (nearly) optimal sample complexity for matrix completion (in terms of number of items, users), and ii) the best known privacy/utility trade-off both theoretically, as well as on benchmark data sets. In particular, we provide the first global convergence analysis of ALS with noise introduced to ensure DP, and show that, in comparison to the best known alternative (the Private Frank-Wolfe algorithm by Jain et al. (2018)), our error bounds scale significantly better with respect to the number of items and users, which is critical in practical problems. Extensive validation on standard benchmarks demonstrate that the algorithm, in combination with carefully designed sampling procedures, is significantly more accurate than existing techniques, thus promising to be the first practical DP embedding model.
****
Light RUMs
Flavio Chierichetti, Ravi Kumar, Andrew Tomkins
A Random Utility Model (RUM) is a distribution on permutations over a universe of items. For each subset of the universe, a RUM induces a natural distribution of the winner in the subset: choose a permutation according to the RUM distribution and pick the maximum item in the subset according to the chosen permutation. RUMs are widely used in the theory of discrete choice. In this paper we consider the question of the (lossy) compressibility of RUMs on a universe of size n, i.e., the minimum number of bits required to approximate the winning probabilities of each slate. Our main result is that RUMs can be approximated using $\tilde{O}(n^2)$ bits, an exponential improvement over the standard representation; fur

thermore, we show that this bound is optimal. En route, we sharpen the classical existential result of McFadden and Train (2000) by showing that the minimum size of a mixture of multinomial logits required to can approximate a general RUM is $\tilde{\Theta}(n)$.
****

Parallelizing Legendre Memory Unit Training

Narsimha Reddy Chilkuri, Chris Eliasmith

Recently, a new recurrent neural network (RNN) named the Legendre Memory Unit (LMU) was proposed and shown to achieve state-of-the-art performance on several benchmark datasets. Here we leverage the linear time-invariant (LTI) memory component of the LMU to construct a simplified variant that can be parallelized during training (and yet executed as an RNN during inference), resulting in up to 200 times faster training. We note that our efficient parallelizing scheme is general and is applicable to any deep network whose recurrent components are linear dynamical systems. We demonstrate the improved accuracy of our new architecture compared to the original LMU and a variety of published LSTM and transformer networks across seven benchmarks. For instance, our LMU sets a new state-of-the-art result on psMNIST, and uses half the parameters while outperforming DistilBERT and LSTM models on IMDB sentiment analysis.
****

Quantifying and Reducing Bias in Maximum Likelihood Estimation of Structured Anomalies

Uthsav Chitra, Kimberly Ding, Jasper C.H. Lee, Benjamin J Raphael

Anomaly estimation, or the problem of finding a subset of a dataset that differs from the rest of the dataset, is a classic problem in machine learning and data mining. In both theoretical work and in applications, the anomaly is assumed to have a specific structure defined by membership in an anomaly family. For example, in temporal data the anomaly family may be time intervals, while in network data the anomaly family may be connected subgraphs. The most prominent approach for anomaly estimation is to compute the Maximum Likelihood Estimator (MLE) of the anomaly; however, it was recently observed that for normally distributed data, the MLE is a biased estimator for some anomaly families. In this work, we demonstrate that in the normal means setting, the bias of the MLE depends on the size of the anomaly family. We prove that if the number of sets in the anomaly family that contain the anomaly is sub-exponential, then the MLE is asymptotically unbiased. We also provide empirical evidence that the converse is true: if the number of such sets is exponential, then the MLE is asymptotically biased. Our analysis unifies a number of earlier results on the bias of the MLE for specific anomaly families. Next, we derive a new anomaly estimator using a mixture model, and we prove that our anomaly estimator is asymptotically unbiased regardless of the size of the anomaly family. We illustrate the advantages of our estimator versus the MLE on disease outbreak data and highway traffic data.
****

Robust Learning-Augmented Caching: An Experimental Study

Jakub Ch■■dowski, Adam Polak, Bartosz Szabucki, Konrad Tomasz ■o■na

Effective caching is crucial for performance of modern-day computing systems. A key optimization problem arising in caching – which item to evict to make room for a new item – cannot be optimally solved without knowing the future. There are many classical approximation algorithms for this problem, but more recently researchers started to successfully apply machine learning to decide what to evict by discovering implicit input patterns and predicting the future. While machine learning typically does not provide any worst-case guarantees, the new field of learning-augmented algorithms proposes solutions which leverage classical online caching algorithms to make the machine-learned predictors robust. We are the first to comprehensively evaluate these learning-augmented algorithms on real-world caching datasets and state-of-the-art machine-learned predictors. We show that a straightforward method – blindly following either a predictor or a classical robust algorithm, and switching whenever one becomes worse than the other – has only a low overhead over a well-performing predictor, while competing with classical methods when the coupled predictor fails, thus providing a cheap worst-case

insurance.
****
Unifying Vision-and-Language Tasks via Text Generation
Jaemin Cho, Jie Lei, Hao Tan, Mohit Bansal
Existing methods for vision-and-language learning typically require designing ta sk-specific architectures and objectives for each task. For example, a multi-lab el answer classifier for visual question answering, a region scorer for referrin g expression comprehension, and a language decoder for image captioning, etc. To alleviate these hassles, in this work, we propose a unified framework that lear ns different tasks in a single architecture with the same language modeling obje ctive, i.e., multimodal conditional text generation, where our models learn to g enerate labels in text based on the visual and textual inputs. On 7 popular visi on-and-language benchmarks, including visual question answering, referring expre ssion comprehension, visual commonsense reasoning, most of which have been previ ously modeled as discriminative tasks, our generative approach (with a single un ified architecture) reaches comparable performance to recent task-specific state -of-the-art vision-and-language models. Moreover, our generative approach shows better generalization ability on questions that have rare answers. Also, we show that our framework allows multi-task learning in a single architecture with a s ingle set of parameters, achieving similar performance to separately optimized s ingle-task models. Our code is publicly available at: https://github.com/j-min/V L-T5
****
Learning from Nested Data with Ornstein Auto-Encoders
Youngwon Choi, Sungdong Lee, Joong-Ho Won
Many of real-world data, e.g., the VGGFace2 dataset, which is a collection of mu ltiple portraits of individuals, come with nested structures due to grouped obse rvation. The Ornstein auto-encoder (OAE) is an emerging framework for representa tion learning from nested data, based on an optimal transport distance between r andom processes. An attractive feature of OAE is its ability to generate new var iations nested within an observational unit, whether or not the unit is known to the model. A previously proposed algorithm for OAE, termed the random-intercept OAE (RIOAE), showed an impressive performance in learning nested representation s, yet lacks theoretical justification. In this work, we show that RIOAE minimiz es a loose upper bound of the employed optimal transport distance. After identif ying several issues with RIOAE, we present the product-space OAE (PSOAE) that mi nimizes a tighter upper bound of the distance and achieves orthogonality in the representation space. PSOAE alleviates the instability of RIOAE and provides mor e flexible representation of nested data. We demonstrate the high performance of PSOAE in the three key tasks of generative models: exemplar generation, style t ransfer, and new concept generation.
****
Variational Empowerment as Representation Learning for Goal-Conditioned Reinforc ement Learning
Jongwook Choi, Archit Sharma, Honglak Lee, Sergey Levine, Shixiang Shane Gu
Learning to reach goal states and learning diverse skills through mutual informa tion maximization have been proposed as principled frameworks for unsupervised r einforcement learning, allowing agents to acquire broadly applicable multi-task policies with minimal reward engineering. In this paper, we discuss how these tw o approaches {—} goal-conditioned RL (GCRL) and MI-based RL {—} can be generaliz ed into a single family of methods, interpreting mutual information maximization and variational empowerment as representation learning methods that acquire fun ction-ally aware state representations for goal reaching.Starting from a simple observation that the standard GCRL is encapsulated by the optimization objective of variational empowerment, we can derive novel variants of GCRL and variationa l empowerment under a single, unified optimization objective, such as adaptive-v ariance GCRL and linear-mapping GCRL, and study the characteristics of represent ation learning each variant provides. Furthermore, through the lens of GCRL, we show that adapting powerful techniques fromGCRL such as goal relabeling into the variationalMI context as well as proper regularization on the variational poste

rior provides substantial gains in algorithm performance, and propose a novel ev
aluation metric named latent goal reaching (LGR)as an objective measure for eval
uating empowerment algorithms akin to goal-based RL. Through principled mathemat
ical derivations and careful experimental validations, our work lays a novel fou
ndation from which representation learning can be evaluated and analyzed in goal
-based RL
****
Label-Only Membership Inference Attacks
Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, Nicolas Paperno
t
Membership inference is one of the simplest privacy threats faced by machine lea
rning models that are trained on private sensitive data. In this attack, an adve
rsary infers whether a particular point was used to train the model, or not, by
observing the model's predictions. Whereas current attack methods all require ac
cess to the model's predicted confidence score, we introduce a label-only attack
 that instead evaluates the robustness of the model's predicted (hard) labels un
der perturbations of the input, to infer membership. Our label-only attack is no
t only as-effective as attacks requiring access to confidence scores, it also de
monstrates that a class of defenses against membership inference, which we call
"confidence masking" because they obfuscate the confidence scores to thwart atta
cks, are insufficient to prevent the leakage of private information. Our experim
ents show that training with differential privacy or strong L2 regularization ar
e the only current defenses that meaningfully decrease leakage of private inform
ation, even for points that are outliers of the training distribution.
****
Modeling Hierarchical Structures with Continuous Recursive Neural Networks
Jishnu Ray Chowdhury, Cornelia Caragea
Recursive Neural Networks (RvNNs), which compose sequences according to their un
derlying hierarchical syntactic structure, have performed well in several natura
l language processing tasks compared to similar models without structural biases
. However, traditional RvNNs are incapable of inducing the latent structure in a
 plain text sequence on their own. Several extensions have been proposed to over
come this limitation. Nevertheless, these extensions tend to rely on surrogate g
radients or reinforcement learning at the cost of higher bias or variance. In th
is work, we propose Continuous Recursive Neural Network (CRvNN) as a backpropaga
tion-friendly alternative to address the aforementioned limitations. This is don
e by incorporating a continuous relaxation to the induced structure. We demonstr
ate that CRvNN achieves strong performance in challenging synthetic tasks such a
s logical inference (Bowman et al., 2015b) and ListOps (Nangia & Bowman, 2018).
We also show that CRvNN performs comparably or better than prior latent structur
e models on real-world tasks such as sentiment analysis and natural language inf
erence.
****
Scaling Multi-Agent Reinforcement Learning with Selective Parameter Sharing
Filippos Christianos, Georgios Papoudakis, Muhammad A Rahman, Stefano V Albrecht
Sharing parameters in multi-agent deep reinforcement learning has played an esse
ntial role in allowing algorithms to scale to a large number of agents. Paramete
r sharing between agents significantly decreases the number of trainable paramet
ers, shortening training times to tractable levels, and has been linked to more
efficient learning. However, having all agents share the same parameters can als
o have a detrimental effect on learning. We demonstrate the impact of parameter
sharing methods on training speed and converged returns, establishing that when
applied indiscriminately, their effectiveness is highly dependent on the environ
ment. We propose a novel method to automatically identify agents which may benef
it from sharing parameters by partitioning them based on their abilities and goa
ls. Our approach combines the increased sample efficiency of parameter sharing w
ith the representational capacity of multiple independent networks to reduce tra
ining time and increase final returns.
****
Beyond Variance Reduction: Understanding the True Impact of Baselines on Policy

Optimization
Wesley Chung, Valentin Thomas, Marlos C. Machado, Nicolas Le Roux
Bandit and reinforcement learning (RL) problems can often be framed as optimization problems where the goal is to maximize average performance while having access only to stochastic estimates of the true gradient. Traditionally, stochastic optimization theory predicts that learning dynamics are governed by the curvature of the loss function and the noise of the gradient estimates. In this paper we demonstrate that the standard view is too limited for bandit and RL problems. To allow our analysis to be interpreted in light of multi-step MDPs, we focus on techniques derived from stochastic optimization principles (e.g., natural policy gradient and EXP3) and we show that some standard assumptions from optimization theory are violated in these problems. We present theoretical results showing that, at least for bandit problems, curvature and noise are not sufficient to explain the learning dynamics and that seemingly innocuous choices like the baseline can determine whether an algorithm converges. These theoretical findings match our empirical evaluation, which we extend to multi-state MDPs.
****
First-Order Methods for Wasserstein Distributionally Robust MDP
Julien Grand Clement, Christian Kroer
Markov decision processes (MDPs) are known to be sensitive to parameter specification. Distributionally robust MDPs alleviate this issue by allowing for \textit{ambiguity sets} which give a set of possible distributions over parameter sets. The goal is to find an optimal policy with respect to the worst-case parameter distribution. We propose a framework for solving Distributionally robust MDPs via first-order methods, and instantiate it for several types of Wasserstein ambiguity sets. By developing efficient proximal updates, our algorithms achieve a convergence rate of $O\left(NA^{2.5}S^{3.5}\log(S)\log(\epsilon^{-1})\epsilon^{-1.5} \right)$ for the number of kernels N in the support of the nominal distribution, states S, and actions A; this rate varies slightly based on the Wasserstein setup. Our dependence on N,A and S is significantly better than existing methods, which have a complexity of $O\left(N^{3.5}A^{3.5}S^{4.5}\log^{2}(\epsilon^{-1}) \right)$. Numerical experiments show that our algorithm is significantly more scalable than state-of-the-art approaches across several domains.
****
Phasic Policy Gradient
Karl W Cobbe, Jacob Hilton, Oleg Klimov, John Schulman
We introduce Phasic Policy Gradient (PPG), a reinforcement learning framework which modifies traditional on-policy actor-critic methods by separating policy and value function training into distinct phases. In prior methods, one must choose between using a shared network or separate networks to represent the policy and value function. Using separate networks avoids interference between objectives, while using a shared network allows useful features to be shared. PPG is able to achieve the best of both worlds by splitting optimization into two phases, one that advances training and one that distills features. PPG also enables the value function to be more aggressively optimized with a higher level of sample reuse. Compared to PPO, we find that PPG significantly improves sample efficiency on the challenging Procgen Benchmark.
****
Riemannian Convex Potential Maps
Samuel Cohen, Brandon Amos, Yaron Lipman
Modeling distributions on Riemannian manifolds is a crucial component in understanding non-Euclidean data that arises, e.g., in physics and geology. The budding approaches in this space are limited by representational and computational tradeoffs. We propose and study a class of flows that uses convex potentials from Riemannian optimal transport. These are universal and can model distributions on any compact Riemannian manifold without requiring domain knowledge of the manifold to be integrated into the architecture. We demonstrate that these flows can model standard distributions on spheres, and tori, on synthetic and geological data.
****

Scaling Properties of Deep Residual Networks
Alain-Sam Cohen, Rama Cont, Alain Rossier, Renyuan Xu
Residual networks (ResNets) have displayed impressive results in pattern recognition and, recently, have garnered considerable theoretical interest due to a perceived link with neural ordinary differential equations (neural ODEs). This link relies on the convergence of network weights to a smooth function as the number of layers increases. We investigate the properties of weights trained by stochastic gradient descent and their scaling with network depth through detailed numerical experiments. We observe the existence of scaling regimes markedly different from those assumed in neural ODE literature. Depending on certain features of the network architecture, such as the smoothness of the activation function, one may obtain an alternative ODE limit, a stochastic differential equation or neither of these. These findings cast doubts on the validity of the neural ODE model as an adequate asymptotic description of deep ResNets and point to an alternative class of differential equations as a better description of the deep network limit.
****
Differentially-Private Clustering of Easy Instances
Edith Cohen, Haim Kaplan, Yishay Mansour, Uri Stemmer, Eliad Tsfadia
Clustering is a fundamental problem in data analysis. In differentially private clustering, the goal is to identify k cluster centers without disclosing information on individual data points. Despite significant research progress, the problem had so far resisted practical solutions. In this work we aim at providing simple implementable differentrially private clustering algorithms when the the data is "easy," e.g., when there exists a significant separation between the clusters. For the easy instances we consider, we have a simple implementation based on utilizing non-private clustering algorithms, and combining them privately. We are able to get improved sample complexity bounds in some cases of Gaussian mixtures and k-means. We complement our theoretical algorithms with experiments of simulated data.
****
Improving Ultrametrics Embeddings Through Coresets
Vincent Cohen-Addad, Rémi De Joannis De Verclos, Guillaume Lagarde
To tackle the curse of dimensionality in data analysis and unsupervised learning, it is critical to be able to efficiently compute “simple” faithful representations of the data that helps extract information, improves understanding and visualization of the structure. When the dataset consists of d-dimensional vectors, simple representations of the data may consist in trees or ultrametrics, and the goal is to best preserve the distances (i.e.: dissimilarity values) between data elements. To circumvent the quadratic running times of the most popular methods for fitting ultrametrics, such as average, single, or complete linkage, \citet{CKL20} recently presented a new algorithm that for any $c \ge 1$, outputs in time $n^{1+O(1/c^2)}$ an ultrametric Δ such that for any two points u, v, $\Delta(u, v)$ is within a multiplicative factor of $5c$ to the distance between u and v in the “best” ultrametric representation. We improve the above result and show how to improve the above guarantee from $5c$ to $\sqrt{2}c + \varepsilon$ while achieving the same asymptotic running time. To complement the improved theoretical bound, we additionally show that the performances of our algorithm are significantly better for various real-world datasets.
****
Correlation Clustering in Constant Many Parallel Rounds
Vincent Cohen-Addad, Silvio Lattanzi, Slobodan Mitrovi■, Ashkan Norouzi-Fard, Nikos Parotsidis, Jakub Tarnawski
Correlation clustering is a central topic in unsupervised learning, with many applications in ML and data mining. In correlation clustering, one receives as input a signed graph and the goal is to partition it to minimize the number of disagreements. In this work we propose a massively parallel computation (MPC) algorithm for this problem that is considerably faster than prior work. In particular, our algorithm uses machines with memory sublinear in the number of nodes in the graph and returns a constant approximation while running only for a constant nu

mber of rounds. To the best of our knowledge, our algorithm is the first that ca
n provably approximate a clustering problem using only a constant number of MPC
rounds in the sublinear memory regime. We complement our analysis with an experi
mental scalability evaluation of our techniques.
****
Concentric mixtures of Mallows models for top-k rankings: sampling and identif
iability
Fabien Collas, Ekhine Irurozki
In this paper, we study mixtures of two Mallows models for top-k rankings with
 equal location parameters but with different scale parameters (a mixture of con
centric Mallows models). These models arise when we have a heterogeneous populat
ion of voters formed by two populations, one of which is a subpopulation of expe
rt voters. We show the identifiability of both components and the learnability o
f their respective parameters. These results are based upon, first, bounding the
 sample complexity for the Borda algorithm with top-k rankings. Second, we cha
racterize the distances between rankings, showing that an off-the-shelf clusteri
ng algorithm separates the rankings by components with high probability -provide
d the scales are well-separated.As a by-product, we include an efficient samplin
g algorithm for Mallows top-k rankings. Finally, since the rank aggregation wi
ll suffer from a large amount of noise introduced by the non-expert voters, we a
dapt the Borda algorithm to be able to recover the ground truth consensus rankin
g which is especially consistent with the expert rankings.
****
Exploiting Shared Representations for Personalized Federated Learning
Liam Collins, Hamed Hassani, Aryan Mokhtari, Sanjay Shakkottai
Deep neural networks have shown the ability to extract universal feature represe
ntations from data such as images and text that have been useful for a variety o
f learning tasks. However, the fruits of representation learning have yet to be
fully-realized in federated settings. Although data in federated settings is oft
en non-i.i.d. across clients, the success of centralized deep learning suggests
that data often shares a global {\em feature representation}, while the statisti
cal heterogeneity across clients or tasks is concentrated in the {\em labels}. B
ased on this intuition, we propose a novel federated learning framework and algo
rithm for learning a shared data representation across clients and unique local
heads for each client. Our algorithm harnesses the distributed computational pow
er across clients to perform many local-updates with respect to the low-dimensio
nal local parameters for every update of the representation. We prove that this
method obtains linear convergence to the ground-truth representation with near-o
ptimal sample complexity in a linear setting, demonstrating that it can efficien
tly reduce the problem dimension for each client. Further, we provide extensive
experimental results demonstrating the improvement of our method over alternativ
e personalized federated learning approaches in heterogeneous settings.
****
Differentiable Particle Filtering via Entropy-Regularized Optimal Transport
Adrien Corenflos, James Thornton, George Deligiannidis, Arnaud Doucet
Particle Filtering (PF) methods are an established class of procedures for perfo
rming inference in non-linear state-space models. Resampling is a key ingredient
 of PF necessary to obtain low variance likelihood and states estimates. However
, traditional resampling methods result in PF-based loss functions being non-dif
ferentiable with respect to model and PF parameters. In a variational inference
context, resampling also yields high variance gradient estimates of the PF-based
 evidence lower bound. By leveraging optimal transport ideas, we introduce a pri
ncipled differentiable particle filter and provide convergence results. We demon
strate this novel method on a variety of applications.
****
Fairness and Bias in Online Selection
Jose Correa, Andres Cristi, Paul Duetting, Ashkan Norouzi-Fard
There is growing awareness and concern about fairness in machine learning and al
gorithm design. This is particularly true in online selection problems where dec
isions are often biased, for example, when assessing credit risks or hiring staf

f. We address the issues of fairness and bias in online selection by introducing multi-color versions of the classic secretary and prophet problem. Interestingly, existing algorithms for these problems are either very unfair or very inefficient, so we develop optimal fair algorithms for these new problems and provide tight bounds on their competitiveness. We validate our theoretical findings on real-world data.
****
Relative Deviation Margin Bounds
Corinna Cortes, Mehryar Mohri, Ananda Theertha Suresh
We present a series of new and more favorable margin-based learning guarantees that depend on the empirical margin loss of a predictor. e give two types of learning bounds, in terms of either the Rademacher complexity or the empirical ℓ_∞-covering number of the hypothesis set used, both distribution-dependent and valid for general families. Furthermore, using our relative deviation margin bounds, we derive distribution-dependent generalization bounds for unbounded loss functions under the assumption of a finite moment. We also briefly highlight several applications of these bounds and discuss their connection with existing results.
****
A Discriminative Technique for Multiple-Source Adaptation
Corinna Cortes, Mehryar Mohri, Ananda Theertha Suresh, Ningshan Zhang
We present a new discriminative technique for the multiple-source adaptation (MSA) problem. Unlike previous work, which relies on density estimation for each source domain, our solution only requires conditional probabilities that can be straightforwardly accurately estimated from unlabeled data from the source domains. We give a detailed analysis of our new technique, including general guarantees based on Rényi divergences, and learning bounds when conditional Maxent is used for estimating conditional probabilities for a point to belong to a source domain. We show that these guarantees compare favorably to those that can be derived for the generative solution, using kernel density estimation. Our experiments with real-world applications further demonstrate that our new discriminative MSA algorithm outperforms the previous generative solution as well as other domain adaptation baselines.
****
Characterizing Fairness Over the Set of Good Models Under Selective Labels
Amanda Coston, Ashesh Rambachan, Alexandra Chouldechova
Algorithmic risk assessments are used to inform decisions in a wide variety of high-stakes settings. Often multiple predictive models deliver similar overall performance but differ markedly in their predictions for individual cases, an empirical phenomenon known as the “Rashomon Effect.” These models may have different properties over various groups, and therefore have different predictive fairness properties. We develop a framework for characterizing predictive fairness properties over the set of models that deliver similar overall performance, or “the set of good models.” Our framework addresses the empirically relevant challenge of selectively labelled data in the setting where the selection decision and outcome are unconfounded given the observed data features. Our framework can be used to 1) audit for predictive bias; or 2) replace an existing model with one that has better fairness properties. We illustrate these use cases on a recidivism prediction task and a real-world credit-scoring task.
****
Two-way kernel matrix puncturing: towards resource-efficient PCA and spectral clustering
Romain Couillet, Florent Chatelain, Nicolas Le Bihan
The article introduces an elementary cost and storage reduction method for spectral clustering and principal component analysis. The method consists in randomly “puncturing” both the data matrix $X\in\mathbb{C}^{p\times n}$ (or $\mathbb{R}^{p\times n}$) and its corresponding kernel (Gram) matrix K through Bernoulli masks: $S\in\{0,1\}^{p\times n}$ for X and $B\in\{0,1\}^{n\times n}$ for K. The resulting “two-way punctured” kernel is thus given by $K=\frac1p[(X\odot S)^\H (X\odot S)]\odot B$. We demonstrate that, for X composed of independent colu

mns drawn from a Gaussian mixture model, as $n,p\to\infty$ with $p/n\to c_0\in(0,\infty)$, the spectral behavior of K – its limiting eigenvalue distribution, as well as its isolated eigenvalues and eigenvectors – is fully tractable and exhibits a series of counter-intuitive phenomena. We notably prove, and empirically confirm on various image databases, that it is possible to drastically puncture the data, thereby providing possibly huge computational and storage gains, for a virtually constant (clustering or PCA) performance. This preliminary study opens as such the path towards rethinking, from a large dimensional standpoint, computational and storage costs in elementary machine learning models.
****

Explaining Time Series Predictions with Dynamic Masks

Jonathan Crabbé, Mihaela Van Der Schaar

How can we explain the predictions of a machine learning model? When the data is structured as a multivariate time series, this question induces additional difficulties such as the necessity for the explanation to embody the time dependency and the large number of inputs. To address these challenges, we propose dynamic masks (Dynamask). This method produces instance-wise importance scores for each feature at each time step by fitting a perturbation mask to the input sequence. In order to incorporate the time dependency of the data, Dynamask studies the effects of dynamic perturbation operators. In order to tackle the large number of inputs, we propose a scheme to make the feature selection parsimonious (to select no more feature than necessary) and legible (a notion that we detail by making a parallel with information theory). With synthetic and real-world data, we demonstrate that the dynamic underpinning of Dynamask, together with its parsimony, offer a neat improvement in the identification of feature importance over time. The modularity of Dynamask makes it ideal as a plug-in to increase the transparency of a wide range of machine learning models in areas such as medicine and finance, where time series are abundant.
****

Generalised Lipschitz Regularisation Equals Distributional Robustness

Zac Cranko, Zhan Shi, Xinhua Zhang, Richard Nock, Simon Kornblith

The problem of adversarial examples has highlighted the need for a theory of regularisation that is general enough to apply to exotic function classes, such as universal approximators. In response, we have been able to significantly sharpen existing results regarding the relationship between distributional robustness and regularisation, when defined with a transportation cost uncertainty set. The theory allows us to characterise the conditions under which the distributional robustness equals a Lipschitz-regularised model, and to tightly quantify, for the first time, the slackness under very mild assumptions. As a theoretical application we show a new result explicating the connection between adversarial learning and distributional robustness. We then give new results for how to achieve Lipschitz regularisation of kernel classifiers, which are demonstrated experimentally.
****

Environment Inference for Invariant Learning

Elliot Creager, Joern-Henrik Jacobsen, Richard Zemel

Learning models that gracefully handle distribution shifts is central to research on domain generalization, robust optimization, and fairness. A promising formulation is domain-invariant learning, which identifies the key issue of learning which features are domain-specific versus domain-invariant. An important assumption in this area is that the training examples are partitioned into “domains” or “environments”. Our focus is on the more common setting where such partitions are not provided. We propose EIIL, a general framework for domain-invariant learning that incorporates Environment Inference to directly infer partitions that are maximally informative for downstream Invariant Learning. We show that EIIL outperforms invariant learning methods on the CMNIST benchmark without using environment labels, and significantly outperforms ERM on worst-group performance in the Waterbirds dataset. Finally, we establish connections between EIIL and algorithmic fairness, which enables EIIL to improve accuracy and calibration in a fair prediction problem.

****
Mind the Box: l_1-APGD for Sparse Adversarial Attacks on Image Classifiers
Francesco Croce, Matthias Hein
We show that when taking into account also the image domain $[0,1]^d$, established l_1-projected gradient descent (PGD) attacks are suboptimal as they do not consider that the effective threat model is the intersection of the l_1-ball and $[0,1]^d$. We study the expected sparsity of the steepest descent step for this effective threat model and show that the exact projection onto this set is computationally feasible and yields better performance. Moreover, we propose an adaptive form of PGD which is highly effective even with a small budget of iterations. Our resulting l_1-APGD is a strong white-box attack showing that prior works overestimated their l_1-robustness. Using l_1-APGD for adversarial training we get a robust classifier with SOTA l_1-robustness. Finally, we combine l_1-APGD and an adaptation of the Square Attack to l_1 into l_1-AutoAttack, an ensemble of attacks which reliably assesses adversarial robustness for the threat model of l_1-ball intersected with $[0,1]^d$.
****
Parameterless Transductive Feature Re-representation for Few-Shot Learning
Wentao Cui, Yuhong Guo
Recent literature in few-shot learning (FSL) has shown that transductive methods often outperform their inductive counterparts. However, most transductive solutions, particularly the meta-learning based ones, require inserting trainable parameters on top of some inductive baselines to facilitate transduction. In this paper, we propose a parameterless transductive feature re-representation framework that differs from all existing solutions from the following perspectives. (1) It is widely compatible with existing FSL methods, including meta-learning and fine tuning based models. (2) The framework is simple and introduces no extra training parameters when applied to any architecture. We conduct experiments on three benchmark datasets by applying the framework to both representative meta-learning baselines and state-of-the-art FSL methods. Our framework consistently improves performances in all experiments and refreshes the state-of-the-art FSL results.
****
Randomized Algorithms for Submodular Function Maximization with a k-System Constraint
Shuang Cui, Kai Han, Tianshuai Zhu, Jing Tang, Benwei Wu, He Huang
Submodular optimization has numerous applications such as crowdsourcing and viral marketing. In this paper, we study the problem of non-negative submodular function maximization subject to a k-system constraint, which generalizes many other important constraints in submodular optimization such as cardinality constraint, matroid constraint, and k-extendible system constraint. The existing approaches for this problem are all based on deterministic algorithmic frameworks, and the best approximation ratio achieved by these algorithms (for a general submodular function) is $k+2\sqrt{k+2}+3$. We propose a randomized algorithm with an improved approximation ratio of $(1+\sqrt{k})^2$, while achieving nearly-linear time complexity significantly lower than that of the state-of-the-art algorithm. We also show that our algorithm can be further generalized to address a stochastic case where the elements can be adaptively selected, and propose an approximation ratio of $(1+\sqrt{k+1})^2$ for the adaptive optimization case. The empirical performance of our algorithms is extensively evaluated in several applications related to data mining and social computing, and the experimental results demonstrate the superiorities of our algorithms in terms of both utility and efficiency.
****
GBHT: Gradient Boosting Histogram Transform for Density Estimation
Jingyi Cui, Hanyuan Hang, Yisen Wang, Zhouchen Lin
In this paper, we propose a density estimation algorithm called \textit{Gradient Boosting Histogram Transform} (GBHT), where we adopt the \textit{Negative Log Likelihood} as the loss function to make the boosting procedure available for the unsupervised tasks. From a learning theory viewpoint, we first prove fast conve

rgence rates for GBHT with the smoothness assumption that the underlying density function lies in the space $C^{0,\alpha}$. Then when the target density function lies in spaces $C^{1,\alpha}$, we present an upper bound for GBHT which is smaller than the lower bound of its corresponding base learner, in the sense of convergence rates. To the best of our knowledge, we make the first attempt to theoretically explain why boosting can enhance the performance of its base learners for density estimation problems. In experiments, we not only conduct performance comparisons with the widely used KDE, but also apply GBHT to anomaly detection to showcase a further application of GBHT.
****
ProGraML: A Graph-based Program Representation for Data Flow Analysis and Compiler Optimizations
Chris Cummins, Zacharias V. Fisches, Tal Ben-Nun, Torsten Hoefler, Michael F P O'Boyle, Hugh Leather
Machine learning (ML) is increasingly seen as a viable approach for building compiler optimization heuristics, but many ML methods cannot replicate even the simplest of the data flow analyses that are critical to making good optimization decisions. We posit that if ML cannot do that, then it is insufficiently able to reason about programs. We formulate data flow analyses as supervised learning tasks and introduce a large open dataset of programs and their corresponding labels from several analyses. We use this dataset to benchmark ML methods and show that they struggle on these fundamental program reasoning tasks. We propose ProGraML - Program Graphs for Machine Learning - a language-independent, portable representation of program semantics. ProGraML overcomes the limitations of prior works and yields improved performance on downstream optimization tasks.
****
Combining Pessimism with Optimism for Robust and Efficient Model-Based Deep Reinforcement Learning
Sebastian Curi, Ilija Bogunovic, Andreas Krause
In real-world tasks, reinforcement learning (RL) agents frequently encounter situations that are not present during training time. To ensure reliable performance, the RL agents need to exhibit robustness to such worst-case situations. The robust-RL framework addresses this challenge via a minimax optimization between an agent and an adversary. Previous robust RL algorithms are either sample inefficient, lack robustness guarantees, or do not scale to large problems. We propose the Robust Hallucinated Upper-Confidence RL (RH-UCRL) algorithm to provably solve this problem while attaining near-optimal sample complexity guarantees. RH-UCRL is a model-based reinforcement learning (MBRL) algorithm that effectively distinguishes between epistemic and aleatoric uncertainty and efficiently explores both the agent and the adversary decision spaces during policy learning. We scale RH-UCRL to complex tasks via neural networks ensemble models as well as neural network policies. Experimentally we demonstrate that RH-UCRL outperforms other robust deep RL algorithms in a variety of adversarial environments.
****
Quantifying Availability and Discovery in Recommender Systems via Stochastic Reachability
Mihaela Curmei, Sarah Dean, Benjamin Recht
In this work, we consider how preference models in interactive recommendation systems determine the availability of content and users' opportunities for discovery. We propose an evaluation procedure based on stochastic reachability to quantify the maximum probability of recommending a target piece of content to an user for a set of allowable strategic modifications. This framework allows us to compute an upper bound on the likelihood of recommendation with minimal assumptions about user behavior. Stochastic reachability can be used to detect biases in the availability of content and diagnose limitations in the opportunities for discovery granted to users. We show that this metric can be computed efficiently as a convex program for a variety of practical settings, and further argue that reachability is not inherently at odds with accuracy. We demonstrate evaluations of recommendation algorithms trained on large datasets of explicit and implicit ratings. Our results illustrate how preference models, selection rules, and user i

nterventions impact reachability and how these effects can be distributed unevenly.
****
Dynamic Balancing for Model Selection in Bandits and RL
Ashok Cutkosky, Christoph Dann, Abhimanyu Das, Claudio Gentile, Aldo Pacchiano, Manish Purohit
We propose a framework for model selection by combining base algorithms in stochastic bandits and reinforcement learning. We require a candidate regret bound for each base algorithm that may or may not hold. We select base algorithms to play in each round using a “balancing condition” on the candidate regret bounds. Our approach simultaneously recovers previous worst-case regret bounds, while also obtaining much smaller regret in natural scenarios when some base learners significantly exceed their candidate bounds. Our framework is relevant in many settings, including linear bandits and MDPs with nested function classes, linear bandits with unknown misspecification, and tuning confidence parameters of algorithms such as LinUCB. Moreover, unlike recent efforts in model selection for linear stochastic bandits, our approach can be extended to consider adversarial rather than stochastic contexts.
****
ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases
Stéphane D’Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, Levent Sagun
Convolutional architectures have proven extremely successful for vision tasks. Their hard inductive biases enable sample-efficient learning, but come at the cost of a potentially lower performance ceiling. Vision Transformers (ViTs) rely on more flexible self-attention layers, and have recently outperformed CNNs for image classification. However, they require costly pre-training on large external datasets or distillation from pre-trained convolutional networks. In this paper, we ask the following question: is it possible to combine the strengths of these two architectures while avoiding their respective limitations? To this end, we introduce gated positional self-attention (GPSA), a form of positional self-attention which can be equipped with a “soft" convolutional inductive bias. We initialise the GPSA layers to mimic the locality of convolutional layers, then give each attention head the freedom to escape locality by adjusting a gating parameter regulating the attention paid to position versus content information. The resulting convolutional-like ViT architecture, ConViT, outperforms the DeiT on ImageNet, while offering a much improved sample efficiency. We further investigate the role of locality in learning by first quantifying how it is encouraged in vanilla self-attention layers, then analysing how it is escaped in GPSA layers. We conclude by presenting various ablations to better understand the success of the ConViT. Our code and models are released publicly at https://github.com/facebookresearch/convit.
****
Consistent regression when oblivious outliers overwhelm
Tommaso D’Orsi, Gleb Novikov, David Steurer
We consider a robust linear regression model $y=X\beta^* + \eta$, where an adversary oblivious to the design $X\in \mathbb{R}^{n\times d}$ may choose η to corrupt all but an α fraction of the observations y in an arbitrary way. Prior to our work, even for Gaussian X, no estimator for β^* was known to be consistent in this model except for quadratic sample size $n \gtrsim (d/\alpha)^2$ or for logarithmic inlier fraction $\alpha\ge 1/\log n$. We show that consistent estimation is possible with nearly linear sample size and inverse-polynomial inlier fraction. Concretely, we show that the Huber loss estimator is consistent for every sample size $n= \omega(d/\alpha^2)$ and achieves an error rate of $O(d/\alpha^2n)^{1/2}$ (both bounds are optimal up to constant factors). Our results extend to designs far beyond the Gaussian case and only require the column span of X to not contain approximately sparse vectors (similar to the kind of assumption commonly made about the kernel space for compressed sensing). We provide two technically similar proofs. One proof is phrased in terms of strong convexity, extending work of [Tsakonas et al. ’14], and particularly short. The

other proof highlights a connection between the Huber loss estimator and high-dimensional median computations. In the special case of Gaussian designs, this connection leads us to a strikingly simple algorithm based on computing coordinate-wise medians that achieves nearly optimal guarantees in linear time, and that can exploit sparsity of β^*. The model studied here also captures heavy-tailed noise distributions that may not even have a first moment.
****
Offline Reinforcement Learning with Pseudometric Learning
Robert Dadashi, Shideh Rezaeifar, Nino Vieillard, Léonard Hussenot, Olivier Pietquin, Matthieu Geist
Offline Reinforcement Learning methods seek to learn a policy from logged transitions of an environment, without any interaction. In the presence of function approximation, and under the assumption of limited coverage of the state-action space of the environment, it is necessary to enforce the policy to visit state-action pairs close to the support of logged transitions. In this work, we propose an iterative procedure to learn a pseudometric (closely related to bisimulation metrics) from logged transitions, and use it to define this notion of closeness. We show its convergence and extend it to the function approximation setting. We then use this pseudometric to define a new lookup based bonus in an actor-critic algorithm: PLOFF. This bonus encourages the actor to stay close, in terms of the defined pseudometric, to the support of logged transitions. Finally, we evaluate the method on hand manipulation and locomotion tasks.
****
A Tale of Two Efficient and Informative Negative Sampling Distributions
Shabnam Daghaghi, Tharun Medini, Nicholas Meisburger, Beidi Chen, Mengnan Zhao, Anshumali Shrivastava
Softmax classifiers with a very large number of classes naturally occur in many applications such as natural language processing and information retrieval. The calculation of full softmax is costly from the computational and energy perspective. There have been various sampling approaches to overcome this challenge, popularly known as negative sampling (NS). Ideally, NS should sample negative classes from a distribution that is dependent on the input data, the current parameters, and the correct positive class. Unfortunately, due to the dynamically updated parameters and data samples, there is no sampling scheme that is provably adaptive and samples the negative classes efficiently. Therefore, alternative heuristics like random sampling, static frequency-based sampling, or learning-based biased sampling, which primarily trade either the sampling cost or the adaptivity of samples per iteration are adopted. In this paper, we show two classes of distributions where the sampling scheme is truly adaptive and provably generates negative samples in near-constant time. Our implementation in C++ on CPU is significantly superior, both in terms of wall-clock time and accuracy, compared to the most optimized TensorFlow implementations of other popular negative sampling approaches on powerful NVIDIA V100 GPU.
****
SiameseXML: Siamese Networks meet Extreme Classifiers with 100M Labels
Kunal Dahiya, Ananye Agarwal, Deepak Saini, Gururaj K, Jian Jiao, Amit Singh, Sumeet Agarwal, Purushottam Kar, Manik Varma
Deep extreme multi-label learning (XML) requires training deep architectures that can tag a data point with its most relevant subset of labels from an extremely large label set. XML applications such as ad and product recommendation involve labels rarely seen during training but which nevertheless hold the key to recommendations that delight users. Effective utilization of label metadata and high quality predictions for rare labels at the scale of millions of labels are thus key challenges in contemporary XML research. To address these, this paper develops the SiameseXML framework based on a novel probabilistic model that naturally motivates a modular approach melding Siamese architectures with high-capacity extreme classifiers, and a training pipeline that effortlessly scales to tasks with 100 million labels. SiameseXML offers predictions 2–13% more accurate than leading XML methods on public benchmark datasets, as well as in live A/B tests on the Bing search engine, it offers significant gains in click-through-rates, cover

age, revenue and other online metrics over state-of-the-art techniques currently in production. Code for SiameseXML is available at https://github.com/Extreme-classification/siamesexml
****
Fixed-Parameter and Approximation Algorithms for PCA with Outliers
Yogesh Dahiya, Fedor Fomin, Fahad Panolan, Kirill Simonov
PCA with Outliers is the fundamental problem of identifying an underlying low-dimensional subspace in a data set corrupted with outliers. A large body of work is devoted to the information-theoretic aspects of this problem. However, from the computational perspective, its complexity is still not well-understood. We study this problem from the perspective of parameterized complexity by investigating how parameters like the dimension of the data, the subspace dimension, the number of outliers and their structure, and approximation error, influence the computational complexity of the problem. Our algorithmic methods are based on techniques of randomized linear algebra and algebraic geometry.
****
Sliced Iterative Normalizing Flows
Biwei Dai, Uros Seljak
We develop an iterative (greedy) deep learning (DL) algorithm which is able to transform an arbitrary probability distribution function (PDF) into the target PDF. The model is based on iterative Optimal Transport of a series of 1D slices, matching on each slice the marginal PDF to the target. The axes of the orthogonal slices are chosen to maximize the PDF difference using Wasserstein distance at each iteration, which enables the algorithm to scale well to high dimensions. As special cases of this algorithm, we introduce two sliced iterative Normalizing Flow (SINF) models, which map from the data to the latent space (GIS) and vice versa (SIG). We show that SIG is able to generate high quality samples of image datasets, which match the GAN benchmarks, while GIS obtains competitive results on density estimation tasks compared to the density trained NFs, and is more stable, faster, and achieves higher p(x) when trained on small training sets. SINF approach deviates significantly from the current DL paradigm, as it is greedy and does not use concepts such as mini-batching, stochastic gradient descent and gradient back-propagation through deep layers.
****
Convex Regularization in Monte-Carlo Tree Search
Tuan Q Dam, Carlo D’Eramo, Jan Peters, Joni Pajarinen
Monte-Carlo planning and Reinforcement Learning (RL) are essential to sequential decision making. The recent AlphaGo and AlphaZero algorithms have shown how to successfully combine these two paradigms to solve large-scale sequential decision problems. These methodologies exploit a variant of the well-known UCT algorithm to trade off the exploitation of good actions and the exploration of unvisited states, but their empirical success comes at the cost of poor sample-efficiency and high computation time. In this paper, we overcome these limitations by introducing the use of convex regularization in Monte-Carlo Tree Search (MCTS) to drive exploration efficiently and to improve policy updates. First, we introduce a unifying theory on the use of generic convex regularizers in MCTS, deriving the first regret analysis of regularized MCTS and showing that it guarantees an exponential convergence rate. Second, we exploit our theoretical framework to introduce novel regularized backup operators for MCTS, based on the relative entropy of the policy update and, more importantly, on the Tsallis entropy of the policy, for which we prove superior theoretical guarantees. We empirically verify the consequence of our theoretical results on a toy problem. Finally, we show how our framework can easily be incorporated in AlphaGo and we empirically show the superiority of convex regularization, w.r.t. representative baselines, on well-known RL problems across several Atari games.
****
Demonstration-Conditioned Reinforcement Learning for Few-Shot Imitation
Christopher R. Dance, Julien Perez, Théo Cachet
In few-shot imitation, an agent is given a few demonstrations of a previously unseen task, and must then successfully perform that task. We propose a novel appr

oach to learning few-shot-imitation agents that we call demonstration-conditione
d reinforcement learning (DCRL). Given a training set consisting of demonstratio
ns, reward functions and transition distributions for multiple tasks, the idea i
s to work with a policy that takes demonstrations as input, and to train this po
licy to maximize the average of the cumulative reward over the set of training t
asks. Relative to previously proposed few-shot imitation methods that use behavi
our cloning or infer reward functions from demonstrations, our method has the di
sadvantage that it requires reward functions at training time. However, DCRL als
o has several advantages, such as the ability to improve upon suboptimal demonst
rations, to operate given state-only demonstrations, and to cope with a domain s
hift between the demonstrator and the agent. Moreover, we show that DCRL outperf
orms methods based on behaviour cloning by a large margin, on navigation tasks a
nd on robotic manipulation tasks from the Meta-World benchmark.
****
Re-understanding Finite-State Representations of Recurrent Policy Networks
Mohamad H Danesh, Anurag Koul, Alan Fern, Saeed Khorram
We introduce an approach for understanding control policies represented as recur
rent neural networks. Recent work has approached this problem by transforming su
ch recurrent policy networks into finite-state machines (FSM) and then analyzing
 the equivalent minimized FSM. While this led to interesting insights, the minim
ization process can obscure a deeper understanding of a machine’s operation by m
erging states that are semantically distinct. To address this issue, we introduc
e an analysis approach that starts with an unminimized FSM and applies more-inte
rpretable reductions that preserve the key decision points of the policy. We als
o contribute an attention tool to attain a deeper understanding of the role of o
bservations in the decisions. Our case studies on 7 Atari games and 3 control be
nchmarks demonstrate that the approach can reveal insights that have not been pr
eviously noticed.
****
Newton Method over Networks is Fast up to the Statistical Precision
Amir Daneshmand, Gesualdo Scutari, Pavel Dvurechensky, Alexander Gasnikov
We propose a distributed cubic regularization of the Newton method for solving (
constrained) empirical risk minimization problems over a network of agents, mode
led as undirected graph. The algorithm employs an inexact, preconditioned Newton
 step at each agent’s side: the gradient of the centralized loss is iteratively
estimated via a gradient-tracking consensus mechanism and the Hessian is subsamp
led over the local data sets. No Hessian matrices are exchanged over the network
. We derive global complexity bounds for convex and strongly convex losses. Our
analysis reveals an interesting interplay between sample and iteration/communica
tion complexity: statistically accurate solutions are achievable in roughly the
same number of iterations of the centralized cubic Newton, with a communication
cost per iteration of the order of $\widetilde{\mathcal{O}}\big(1/\sqrt{1-\rho}\big)$, where ρ characterizes the connectivity of the network. This represen
ts a significant improvement with respect to existing, statistically oblivious,
distributed Newton-based methods over networks.
****
BasisDeVAE: Interpretable Simultaneous Dimensionality Reduction and Feature-Leve
l Clustering with Derivative-Based Variational Autoencoders
Dominic Danks, Christopher Yau
The Variational Autoencoder (VAE) performs effective nonlinear dimensionality re
duction in a variety of problem settings. However, the black-box neural network
decoder function typically employed limits the ability of the decoder function t
o be constrained and interpreted, making the use of VAEs problematic in settings
 where prior knowledge should be embedded within the decoder. We present DeVAE,
a novel VAE-based model with a derivative-based forward mapping, allowing for gr
eater control over decoder behaviour via specification of the decoder function i
n derivative space. Additionally, we show how DeVAE can be paired with a sparse
clustering prior to create BasisDeVAE and perform interpretable simultaneous dim
ensionality reduction and feature-level clustering. We demonstrate the performan
ce and scalability of the DeVAE and BasisDeVAE models on synthetic and real-worl

d data and present how the derivative-based approach allows for expressive yet interpretable forward models which respect prior knowledge.
****
Intermediate Layer Optimization for Inverse Problems using Deep Generative Models
Giannis Daras, Joseph Dean, Ajil Jalal, Alex Dimakis
We propose Intermediate Layer Optimization (ILO), a novel optimization algorithm for solving inverse problems with deep generative models. Instead of optimizing only over the initial latent code, we progressively change the input layer obtaining successively more expressive generators. To explore the higher dimensional spaces, our method searches for latent codes that lie within a small l1 ball around the manifold induced by the previous layer. Our theoretical analysis shows that by keeping the radius of the ball relatively small, we can improve the established error bound for compressed sensing with deep generative models. We empirically show that our approach outperforms state-of-the-art methods introduced in StyleGAN2 and PULSE for a wide range of inverse problems including inpainting, denoising, super-resolution and compressed sensing.
****
Measuring Robustness in Deep Learning Based Compressive Sensing
Mohammad Zalbagi Darestani, Akshay S Chaudhari, Reinhard Heckel
Deep neural networks give state-of-the-art accuracy for reconstructing images from few and noisy measurements, a problem arising for example in accelerated magnetic resonance imaging (MRI). However, recent works have raised concerns that deep-learning-based image reconstruction methods are sensitive to perturbations and are less robust than traditional methods: Neural networks (i) may be sensitive to small, yet adversarially-selected perturbations, (ii) may perform poorly under distribution shifts, and (iii) may fail to recover small but important features in an image. In order to understand the sensitivity to such perturbations, in this work, we measure the robustness of different approaches for image reconstruction including trained and un-trained neural networks as well as traditional sparsity-based methods. We find, contrary to prior works, that both trained and un-trained methods are vulnerable to adversarial perturbations. Moreover, both trained and un-trained methods tuned for a particular dataset suffer very similarly from distribution shifts. Finally, we demonstrate that an image reconstruction method that achieves higher reconstruction quality, also performs better in terms of accurately recovering fine details. Our results indicate that the state-of-the-art deep-learning-based image reconstruction methods provide improved performance than traditional methods without compromising robustness.
****
SAINT-ACC: Safety-Aware Intelligent Adaptive Cruise Control for Autonomous Vehicles Using Deep Reinforcement Learning
Lokesh Chandra Das, Myounggyu Won
We present a novel adaptive cruise control (ACC) system namely SAINT-ACC: {S}afety-{A}ware {Int}elligent {ACC} system (SAINT-ACC) that is designed to achieve simultaneous optimization of traffic efficiency, driving safety, and driving comfort through dynamic adaptation of the inter-vehicle gap based on deep reinforcement learning (RL). A novel dual RL agent-based approach is developed to seek and adapt the optimal balance between traffic efficiency and driving safety/comfort by effectively controlling the driving safety model parameters and inter-vehicle gap based on macroscopic and microscopic traffic information collected from dynamically changing and complex traffic environments. Results obtained through over 12,000 simulation runs with varying traffic scenarios and penetration rates demonstrate that SAINT-ACC significantly enhances traffic flow, driving safety and comfort compared with a state-of-the-art approach.
****
Lipschitz normalization for self-attention layers with application to graph neural networks
George Dasoulas, Kevin Scaman, Aladin Virmaux
Attention based neural networks are state of the art in a large range of applications. However, their performance tends to degrade when the number of layers inc

reases. In this work, we show that enforcing Lipschitz continuity by normalizing
 the attention scores can significantly improve the performance of deep attentio
n models. First, we show that, for deep graph attention networks (GAT), gradient
 explosion appears during training, leading to poor performance of gradient-base
d training algorithms. To address this issue, we derive a theoretical analysis o
f the Lipschitz continuity of attention modules and introduce LipschitzNorm, a s
imple and parameter-free normalization for self-attention mechanisms that enforc
es the model to be Lipschitz continuous. We then apply LipschitzNorm to GAT and
Graph Transformers and show that their performance is substantially improved in
the deep setting (10 to 30 layers). More specifically, we show that a deep GAT m
odel with LipschitzNorm achieves state of the art results for node label predict
ion tasks that exhibit long-range dependencies, while showing consistent improve
ments over their unnormalized counterparts in benchmark node classification task
s.
****
Householder Sketch for Accurate and Accelerated Least-Mean-Squares Solvers
Jyotikrishna Dass, Rabi Mahapatra
Least-Mean-Squares (\textsc{LMS}) solvers comprise a class of fundamental optimi
zation problems such as linear regression, and regularized regressions such as R
idge, LASSO, and Elastic-Net. Data summarization techniques for big data generat
e summaries called coresets and sketches to speed up model learning under stream
ing and distributed settings. For example, \citep{nips2019} design a fast and ac
curate Caratheodory set on input data to boost the performance of existing \text
sc{LMS} solvers. In retrospect, we explore classical Householder transformation
as a candidate for sketching and accurately solving LMS problems. We find it to
be a simpler, memory-efficient, and faster alternative that always existed to th
e above strong baseline. We also present a scalable algorithm based on the const
ruction of distributed Householder sketches to solve \textsc{LMS} problem across
 multiple worker nodes. We perform thorough empirical analysis with large synthe
tic and real datasets to evaluate the performance of Householder sketch and comp
are with \citep{nips2019}. Our results show Householder sketch speeds up existin
g \textsc{LMS} solvers in the scikit-learn library up to 100x-400x. Also, it
 is 10x-100x faster than the above baseline with similar numerical stability
. The distributed algorithm demonstrates linear scalability with a near-negligib
le communication overhead.
****
Byzantine-Resilient High-Dimensional SGD with Local Iterations on Heterogeneous
Data
Deepesh Data, Suhas Diggavi
We study stochastic gradient descent (SGD) with local iterations in the presence
 of Byzantine clients, motivated by the federated learning. The clients, instead
 of communicating with the server in every iteration, maintain their local model
s, which they update by taking several SGD iterations based on their own dataset
s and then communicate the net update with the server, thereby achieving communi
cation-efficiency. Furthermore, only a subset of clients communicates with the s
erver at synchronization times. The Byzantine clients may collude and send arbit
rary vectors to the server to disrupt the learning process. To combat the advers
ary, we employ an efficient high-dimensional robust mean estimation algorithm at
 the server to filter-out corrupt vectors; and to analyze the outlier-filtering
procedure, we develop a novel matrix concentration result that may be of indepen
dent interest. We provide convergence analyses for both strongly-convex and non-
convex smooth objectives in the heterogeneous data setting. We believe that ours
 is the first Byzantine-resilient local SGD algorithm and analysis with non-triv
ial guarantees. We corroborate our theoretical results with preliminary experime
nts for neural network training.
****
Catformer: Designing Stable Transformers via Sensitivity Analysis
Jared Q Davis, Albert Gu, Krzysztof Choromanski, Tri Dao, Christopher Re, Chelse
a Finn, Percy Liang
Transformer architectures are widely used, but training them is non-trivial, req

uiring custom learning rate schedules, scaling terms, residual connections, careful placement of submodules such as normalization, and so on. In this paper, we improve upon recent analysis of Transformers and formalize a notion of sensitivity to capture the difficulty of training. Sensitivity characterizes how the variance of activation and gradient norms change in expectation when parameters are randomly perturbed. We analyze the sensitivity of previous Transformer architectures and design a new architecture, the Catformer, which replaces residual connections or RNN-based gating mechanisms with concatenation. We prove that Catformers are less sensitive than other Transformer variants and demonstrate that this leads to more stable training. On DMLab30, a suite of high-dimension reinforcement tasks, Catformer outperforms other transformers, including Gated Transformer-XL—the state-of-the-art architecture designed to address stability—by 13%.
****
Diffusion Source Identification on Networks with Statistical Confidence
Quinlan E Dawkins, Tianxi Li, Haifeng Xu
Diffusion source identification on networks is a problem of fundamental importance in a broad class of applications, including controlling the spreading of rumors on social media, identifying a computer virus over cyber networks, or identifying the disease center during epidemiology. Though this problem has received significant recent attention, most known approaches are well-studied in only very restrictive settings and lack theoretical guarantees for more realistic networks. We introduce a statistical framework for the study of this problem and develop a confidence set inference approach inspired by hypothesis testing. Our method efficiently produces a small subset of nodes, which provably covers the source node with any pre-specified confidence level without restrictive assumptions on network structures. To our knowledge, this is the first diffusion source identification method with a practically useful theoretical guarantee on general networks. We demonstrate our approach via extensive synthetic experiments on well-known random network models, a large data set of real-world networks as well as a mobility network between cities concerning the COVID-19 spreading in January 2020.
****
Bayesian Deep Learning via Subnetwork Inference
Erik Daxberger, Eric Nalisnick, James U Allingham, Javier Antoran, Jose Miguel Hernandez-Lobato
The Bayesian paradigm has the potential to solve core issues of deep neural networks such as poor calibration and data inefficiency. Alas, scaling Bayesian inference to large weight spaces often requires restrictive approximations. In this work, we show that it suffices to perform inference over a small subset of model weights in order to obtain accurate predictive posteriors. The other weights are kept as point estimates. This subnetwork inference framework enables us to use expressive, otherwise intractable, posterior approximations over such subsets. In particular, we implement subnetwork linearized Laplace as a simple, scalable Bayesian deep learning method: We first obtain a MAP estimate of all weights and then infer a full-covariance Gaussian posterior over a subnetwork using the linearized Laplace approximation. We propose a subnetwork selection strategy that aims to maximally preserve the model’s predictive uncertainty. Empirically, our approach compares favorably to ensembles and less expressive posterior approximations over full networks.
****
Adversarial Robustness Guarantees for Random Deep Neural Networks
Giacomo De Palma, Bobak Kiani, Seth Lloyd
The reliability of deep learning algorithms is fundamentally challenged by the existence of adversarial examples, which are incorrectly classified inputs that are extremely close to a correctly classified input. We explore the properties of adversarial examples for deep neural networks with random weights and biases, and prove that for any p$\geq$1, the \ell^p distance of any given input from the classification boundary scales as one over the square root of the dimension of the input times the \ell^p norm of the input. The results are based on the recently proved equivalence between Gaussian processes and deep neural networks in the limit of infinite width of the hidden layers, and are validated with experiment

s on both random deep neural networks and deep neural networks trained on the MN
IST and CIFAR10 datasets. The results constitute a fundamental advance in the th
eoretical understanding of adversarial examples, and open the way to a thorough
theoretical characterization of the relation between network architecture and ro
bustness to adversarial perturbations.
****
High-Dimensional Gaussian Process Inference with Derivatives
Filip de Roos, Alexandra Gessner, Philipp Hennig
Although it is widely known that Gaussian processes can be conditioned on observ
ations of the gradient, this functionality is of limited use due to the prohibit
ive computational cost of $\mathcal{O}(N^3 D^3)$ in data points N and dimensio
n D. The dilemma of gradient observations is that a single one of them comes a
t the same cost as D independent function evaluations, so the latter are often
 preferred. Careful scrutiny reveals, however, that derivative observations give
 rise to highly structured kernel Gram matrices for very general classes of kern
els (inter alia, stationary kernels). We show that in the \emph{low-data} regime
 $N < D$, the Gram matrix can be decomposed in a manner that reduces the cost of
 inference to $\mathcal{O}(N^2D + (N^2)^3)$ (i.e., linear in the number of dimen
sions) and, in special cases, to $\mathcal{O}(N^2D + N^3)$. This reduction in co
mplexity opens up new use-cases for inference with gradients especially in the h
igh-dimensional regime, where the information-to-cost ratio of gradient observat
ions significantly increases. We demonstrate this potential in a variety of task
s relevant for machine learning, such as optimization and Hamiltonian Monte Carl
o with predictive gradients.
****
Transfer-Based Semantic Anomaly Detection
Lucas Deecke, Lukas Ruff, Robert A. Vandermeulen, Hakan Bilen
Detecting semantic anomalies is challenging due to the countless ways in which t
hey may appear in real-world data. While enhancing the robustness of networks ma
y be sufficient for modeling simplistic anomalies, there is no good known way of
 preparing models for all potential and unseen anomalies that can potentially oc
cur, such as the appearance of new object classes. In this paper, we show that a
 previously overlooked strategy for anomaly detection (AD) is to introduce an ex
plicit inductive bias toward representations transferred over from some large an
d varied semantic task. We rigorously verify our hypothesis in controlled trials
 that utilize intervention, and show that it gives rise to surprisingly effectiv
e auxiliary objectives that outperform previous AD paradigms.
****
Grid-Functioned Neural Networks
Javier Dehesa, Andrew Vidler, Julian Padget, Christof Lutteroth
We introduce a new neural network architecture that we call "grid-functioned" ne
ural networks. It utilises a grid structure of network parameterisations that ca
n be specialised for different subdomains of the problem, while maintaining smoo
th, continuous behaviour. The grid gives the user flexibility to prevent gross f
eatures from overshadowing important minor ones. We present a full characterisat
ion of its computational and spatial complexity, and demonstrate its potential,
compared to a traditional architecture, over a set of synthetic regression probl
ems. We further illustrate the benefits through a real-world 3D skeletal animati
on case study, where it offers the same visual quality as a state-of-the-art mod
el, but with lower computational complexity and better control accuracy.
****
Multidimensional Scaling: Approximation and Complexity
Erik Demaine, Adam Hesterberg, Frederic Koehler, Jayson Lynch, John Urschel
Metric Multidimensional scaling (MDS) is a classical method for generating meani
ngful (non-linear) low-dimensional embeddings of high-dimensional data. MDS has
a long history in the statistics, machine learning, and graph drawing communitie
s. In particular, the Kamada-Kawai force-directed graph drawing method is equiva
lent to MDS and is one of the most popular ways in practice to embed graphs into
 low dimensions. Despite its ubiquity, our theoretical understanding of MDS rema
ins limited as its objective function is highly non-convex. In this paper, we pr

ove that minimizing the Kamada-Kawai objective is NP-hard and give a provable ap proximation algorithm for optimizing it, which in particular is a PTAS on low-di ameter graphs. We supplement this result with experiments suggesting possible co nnections between our greedy approximation algorithm and gradient-based methods.
****
What Does Rotation Prediction Tell Us about Classifier Accuracy under Varying Te sting Environments?
Weijian Deng, Stephen Gould, Liang Zheng
Understanding classifier decision under novel environments is central to the com munity, and a common practice is evaluating it on labeled test sets. However, in real-world testing, image annotations are difficult and expensive to obtain, es pecially when the test environment is changing. A natural question then arises: given a trained classifier, can we evaluate its accuracy on varying unlabeled te st sets? In this work, we train semantic classification and rotation prediction in a multi-task way. On a series of datasets, we report an interesting finding, i.e., the semantic classification accuracy exhibits a strong linear relationship with the accuracy of the rotation prediction task (Pearson’s Correlation r > 0. 88). This finding allows us to utilize linear regression to estimate classifier performance from the accuracy of rotation prediction which can be obtained on th e test set through the freely generated rotation labels.
****
Toward Better Generalization Bounds with Locally Elastic Stability
Zhun Deng, Hangfeng He, Weijie Su
Algorithmic stability is a key characteristic to ensure the generalization abili ty of a learning algorithm. Among different notions of stability, \emph{uniform stability} is arguably the most popular one, which yields exponential generaliza tion bounds. However, uniform stability only considers the worst-case loss chang e (or so-called sensitivity) by removing a single data point, which is distribut ion-independent and therefore undesirable. There are many cases that the worst-c ase sensitivity of the loss is much larger than the average sensitivity taken ov er the single data point that is removed, especially in some advanced models suc h as random feature models or neural networks. Many previous works try to mitiga te the distribution independent issue by proposing weaker notions of stability, however, they either only yield polynomial bounds or the bounds derived do not v anish as sample size goes to infinity. Given that, we propose \emph{locally elas tic stability} as a weaker and distribution-dependent stability notion, which st ill yields exponential generalization bounds. We further demonstrate that locall y elastic stability implies tighter generalization bounds than those derived bas ed on uniform stability in many situations by revisiting the examples of bounded support vector machines, regularized least square regressions, and stochastic g radient descent.
****
Revenue-Incentive Tradeoffs in Dynamic Reserve Pricing
Yuan Deng, Sebastien Lahaie, Vahab Mirrokni, Song Zuo
Online advertisements are primarily sold via repeated auctions with reserve pric es. In this paper, we study how to set reserves to boost revenue based on the hi storical bids of strategic buyers, while controlling the impact of such a policy on the incentive compatibility of the repeated auctions. Adopting an incentive compatibility metric which quantifies the incentives to shade bids, we propose a novel class of reserve pricing policies and provide analytical tradeoffs betwee n their revenue performance and bid-shading incentives. The policies are inspire d by the exponential mechanism from the literature on differential privacy, but our study uncovers mechanisms with significantly better revenue-incentive tradeo ffs than the exponential mechanism in practice. We further empirically evaluate the tradeoffs on synthetic data as well as real ad auction data from a major ad exchange to verify and support our theoretical findings.
****
Heterogeneity for the Win: One-Shot Federated Clustering
Don Kurian Dennis, Tian Li, Virginia Smith
In this work, we explore the unique challenges—and opportunities—of unsupervised

federated learning (FL). We develop and analyze a one-shot federated clustering scheme, kfed, based on the widely-used Lloyd's method for k-means clustering. In contrast to many supervised problems, we show that the issue of statistical heterogeneity in federated networks can in fact benefit our analysis. We analyse kfed under a center separation assumption and compare it to the best known requirements of its centralized counterpart. Our analysis shows that in heterogeneous regimes where the number of clusters per device (k') is smaller than the total number of clusters over the network k, $(k'\le \sqrt{k})$, we can use heterogeneity to our advantage—significantly weakening the cluster separation requirements for kfed. From a practical viewpoint, kfed also has many desirable properties: it requires only round of communication, can run asynchronously, and can handle partial participation or node/network failures. We motivate our analysis with experiments on common FL benchmarks, and highlight the practical utility of one-shot clustering through use-cases in personalized FL and device sampling.
****

Kernel Continual Learning
Mohammad Mahdi Derakhshani, Xiantong Zhen, Ling Shao, Cees Snoek
This paper introduces kernel continual learning, a simple but effective variant of continual learning that leverages the non-parametric nature of kernel methods to tackle catastrophic forgetting. We deploy an episodic memory unit that stores a subset of samples for each task to learn task-specific classifiers based on kernel ridge regression. This does not require memory replay and systematically avoids task interference in the classifiers. We further introduce variational random features to learn a data-driven kernel for each task. To do so, we formulate kernel continual learning as a variational inference problem, where a random Fourier basis is incorporated as the latent variable. The variational posterior distribution over the random Fourier basis is inferred from the coreset of each task. In this way, we are able to generate more informative kernels specific to each task, and, more importantly, the coreset size can be reduced to achieve more compact memory, resulting in more efficient continual learning based on episodic memory. Extensive evaluation on four benchmarks demonstrates the effectiveness and promise of kernels for continual learning.
****

Bayesian Optimization over Hybrid Spaces
Aryan Deshwal, Syrine Belakaria, Janardhan Rao Doppa
We consider the problem of optimizing hybrid structures (mixture of discrete and continuous input variables) via expensive black-box function evaluations. This problem arises in many real-world applications. For example, in materials design optimization via lab experiments, discrete and continuous variables correspond to the presence/absence of primitive elements and their relative concentrations respectively. The key challenge is to accurately model the complex interactions between discrete and continuous variables. In this paper, we propose a novel approach referred as Hybrid Bayesian Optimization (HyBO) by utilizing diffusion kernels, which are naturally defined over continuous and discrete variables. We develop a principled approach for constructing diffusion kernels over hybrid spaces by utilizing the additive kernel formulation, which allows additive interactions of all orders in a tractable manner. We theoretically analyze the modeling strength of additive hybrid kernels and prove that it has the universal approximation property. Our experiments on synthetic and six diverse real-world benchmarks show that HyBO significantly outperforms the state-of-the-art methods.
****

Navigation Turing Test (NTT): Learning to Evaluate Human-Like Navigation
Sam Devlin, Raluca Georgescu, Ida Momennejad, Jaroslaw Rzepecki, Evelyn Zuniga, Gavin Costello, Guy Leroy, Ali Shaw, Katja Hofmann
A key challenge on the path to developing agents that learn complex human-like behavior is the need to quickly and accurately quantify human-likeness. While human assessments of such behavior can be highly accurate, speed and scalability are limited. We address these limitations through a novel automated Navigation Turing Test (ANTT) that learns to predict human judgments of human-likeness. We demonstrate the effectiveness of our automated NTT on a navigation task in a comple

x 3D environment. We investigate six classification models to shed light on the
types of architectures best suited to this task, and validate them against data
collected through a human NTT. Our best models achieve high accuracy when distin
guishing true human and agent behavior. At the same time, we show that predictin
g finer-grained human assessment of agents' progress towards human-like behavior
 remains unsolved. Our work takes an important step towards agents that more eff
ectively learn complex human-like behavior.
****
Versatile Verification of Tree Ensembles
Laurens Devos, Wannes Meert, Jesse Davis
Machine learned models often must abide by certain requirements (e.g., fairness
or legal). This has spurred interested in developing approaches that can provabl
y verify whether a model satisfies certain properties. This paper introduces a g
eneric algorithm called Veritas that enables tackling multiple different verific
ation tasks for tree ensemble models like random forests (RFs) and gradient boos
ted decision trees (GBDTs). This generality contrasts with previous work, which
has focused exclusively on either adversarial example generation or robustness c
hecking. Veritas formulates the verification task as a generic optimization prob
lem and introduces a novel search space representation. Veritas offers two key a
dvantages. First, it provides anytime lower and upper bounds when the optimizati
on problem cannot be solved exactly. In contrast, many existing methods have foc
used on exact solutions and are thus limited by the verification problem being N
P-complete. Second, Veritas produces full (bounded suboptimal) solutions that ca
n be used to generate concrete examples. We experimentally show that our method
produces state-of-the-art robustness estimates, especially when executed with st
rict time constraints. This is exceedingly important when checking the robustnes
s of large datasets. Additionally, we show that Veritas enables tackling more re
al-world verification scenarios.
****
On the Inherent Regularization Effects of Noise Injection During Training
Oussama Dhifallah, Yue Lu
Randomly perturbing networks during the training process is a commonly used appr
oach to improving generalization performance. In this paper, we present a theore
tical study of one particular way of random perturbation, which corresponds to i
njecting artificial noise to the training data. We provide a precise asymptotic
characterization of the training and generalization errors of such randomly pert
urbed learning problems on a random feature model. Our analysis shows that Gauss
ian noise injection in the training process is equivalent to introducing a weigh
ted ridge regularization, when the number of noise injections tends to infinity.
 The explicit form of the regularization is also given. Numerical results corrob
orate our asymptotic predictions, showing that they are accurate even in moderat
e problem dimensions. Our theoretical predictions are based on a new correlated
Gaussian equivalence conjecture that generalizes recent results in the study of
random feature models.
****
Hierarchical Agglomerative Graph Clustering in Nearly-Linear Time
Laxman Dhulipala, David Eisenstat, Jakub ■■cki, Vahab Mirrokni, Jessica Shi
We study the widely-used hierarchical agglomerative clustering (HAC) algorithm o
n edge-weighted graphs. We define an algorithmic framework for hierarchical aggl
omerative graph clustering that provides the first efficient $\tilde{O}(m)$ time
 exact algorithms for classic linkage measures, such as complete- and WPGMA-link
age, as well as other measures. Furthermore, for average-linkage, arguably the m
ost popular variant of HAC, we provide an algorithm that runs in $\tilde{O}(n\sq
rt{m})$ time. For this variant, this is the first exact algorithm that runs in s
ubquadratic time, as long as $m=n^{2-\epsilon}$ for some constant $\epsilon > 0$
. We complement this result with a simple ϵ-close approximation algorit
hm for average-linkage in our framework that runs in $\tilde{O}(m)$ time. As an
application of our algorithms, we consider clustering points in a metric space b
y first using k-NN to generate a graph from the point set, and then running ou
r algorithms on the resulting weighted graph. We validate the performance of our

algorithms on publicly available datasets, and show that our approach can speed up clustering of point datasets by a factor of 20.7–76.5x.
****
Learning Online Algorithms with Distributional Advice
Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, Ali Vakilian, Nikos Zarifis
We study the problem of designing online algorithms given advice about the input. While prior work had focused on deterministic advice, we only assume distributional access to the instances of interest, and the goal is to learn a competitive algorithm given access to i.i.d. samples. We aim to be competitive against an adversary with prior knowledge of the distribution, while also performing well against worst-case inputs. We focus on the classical online problems of ski-rental and prophet-inequalities, and provide sample complexity bounds for the underlying learning tasks. First, we point out that for general distributions it is information-theoretically impossible to beat the worst-case competitive-ratio with any finite sample size. As our main contribution, we establish strong positive results for well-behaved distributions. Specifically, for the broad class of log-concave distributions, we show that $\mathrm{poly}(1/\epsilon)$ samples suffice to obtain $(1+\epsilon)$-competitive ratio. Finally, we show that this sample upper bound is close to best possible, even for very simple classes of distributions.
****
A Wasserstein Minimax Framework for Mixed Linear Regression
Theo Diamandis, Yonina Eldar, Alireza Fallah, Farzan Farnia, Asuman Ozdaglar
Multi-modal distributions are commonly used to model clustered data in statistical learning tasks. In this paper, we consider the Mixed Linear Regression (MLR) problem. We propose an optimal transport-based framework for MLR problems, Wasserstein Mixed Linear Regression (WMLR), which minimizes the Wasserstein distance between the learned and target mixture regression models. Through a model-based duality analysis, WMLR reduces the underlying MLR task to a nonconvex-concave minimax optimization problem, which can be provably solved to find a minimax stationary point by the Gradient Descent Ascent (GDA) algorithm. In the special case of mixtures of two linear regression models, we show that WMLR enjoys global convergence and generalization guarantees. We prove that WMLR’s sample complexity grows linearly with the dimension of data. Finally, we discuss the application of WMLR to the federated learning task where the training samples are collected by multiple agents in a network. Unlike the Expectation-Maximization algorithm, WMLR directly extends to the distributed, federated learning setting. We support our theoretical results through several numerical experiments, which highlight our framework’s ability to handle the federated learning setting with mixture models.
****
Context-Aware Online Collective Inference for Templated Graphical Models
Charles Dickens, Connor Pryor, Eriq Augustine, Alexander Miller, Lise Getoor
In this work, we examine online collective inference, the problem of maintaining and performing inference over a sequence of evolving graphical models. We utilize templated graphical models (TGM), a general class of graphical models expressed via templates and instantiated with data. A key challenge is minimizing the cost of instantiating the updated model. To address this, we define a class of exact and approximate context-aware methods for updating an existing TGM. These methods avoid a full re-instantiation by using the context of the updates to only add relevant components to the graphical model. Further, we provide stability bounds for the general online inference problem and regret bounds for a proposed approximation. Finally, we implement our approach in probabilistic soft logic, and test it on several online collective inference tasks. Through these experiments we verify the bounds on regret and stability, and show that our approximate online approach consistently runs two to five times faster than the offline alternative while, surprisingly, maintaining the quality of the predictions.
****
ARMS: Antithetic-REINFORCE-Multi-Sample Gradient for Binary Variables

Aleksandar Dimitriev, Mingyuan Zhou

Estimating the gradients for binary variables is a task that arises frequently in various domains, such as training discrete latent variable models. What has been commonly used is a REINFORCE based Monte Carlo estimation method that uses either independent samples or pairs of negatively correlated samples. To better utilize more than two samples, we propose ARMS, an Antithetic REINFORCE-based Multi-Sample gradient estimator. ARMS uses a copula to generate any number of mutually antithetic samples. It is unbiased, has low variance, and generalizes both DisARM, which we show to be ARMS with two samples, and the leave-one-out REINFORCE (LOORF) estimator, which is ARMS with uncorrelated samples. We evaluate ARMS on several datasets for training generative models, and our experimental results show that it outperforms competing methods. We also develop a version of ARMS for optimizing the multi-sample variational bound, and show that it outperforms both VIMCO and DisARM. The code is publicly available.

****

XOR-CD: Linearly Convergent Constrained Structure Generation

Fan Ding, Jianzhu Ma, Jinbo Xu, Yexiang Xue

We propose XOR-Contrastive Divergence learning (XOR-CD), a provable approach for constrained structure generation, which remains difficult for state-of-the-art neural network and constraint reasoning approaches. XOR-CD harnesses XOR-Sampling to generate samples from the model distribution in CD learning and is guaranteed to generate valid structures. In addition, XOR-CD has a linear convergence rate towards the global maximum of the likelihood function within a vanishing constant in learning exponential family models. Constraint satisfaction enabled by XOR-CD also boosts its learning performance. Our real-world experiments on data-driven experimental design, dispatching route generation, and sequence-based protein homology detection demonstrate the superior performance of XOR-CD compared to baseline approaches in generating valid structures as well as capturing the inductive bias in the training set.

****

Dual Principal Component Pursuit for Robust Subspace Learning: Theory and Algorithms for a Holistic Approach

Tianyu Ding, Zhihui Zhu, Rene Vidal, Daniel P Robinson

The Dual Principal Component Pursuit (DPCP) method has been proposed to robustly recover a subspace of high-relative dimension from corrupted data. Existing analyses and algorithms of DPCP, however, mainly focus on finding a normal to a single hyperplane that contains the inliers. Although these algorithms can be extended to a subspace of higher co-dimension through a recursive approach that sequentially finds a new basis element of the space orthogonal to the subspace, this procedure is computationally expensive and lacks convergence guarantees. In this paper, we consider a DPCP approach for simultaneously computing the entire basis of the orthogonal complement subspace (we call this a holistic approach) by solving a non-convex non-smooth optimization problem over the Grassmannian. We provide geometric and statistical analyses for the global optimality and prove that it can tolerate as many outliers as the square of the number of inliers, under both noiseless and noisy settings. We then present a Riemannian regularity condition for the problem, which is then used to prove that a Riemannian subgradient method converges linearly to a neighborhood of the orthogonal subspace with error proportional to the noise level.

****

Coded-InvNet for Resilient Prediction Serving Systems

Tuan Dinh, Kangwook Lee

Inspired by a new coded computation algorithm for invertible functions, we propose Coded-InvNet a new approach to design resilient prediction serving systems that can gracefully handle stragglers or node failures. Coded-InvNet leverages recent findings in the deep learning literature such as invertible neural networks, Manifold Mixup, and domain translation algorithms, identifying interesting research directions that span across machine learning and systems. Our experimental results show that Coded-InvNet can outperform existing approaches, especially when the compute resource overhead is as low as 10%. For instance, without knowing

which of the ten workers is going to fail, our algorithm can design a backup task so that it can correctly recover the missing prediction result with an accuracy of 85.9%, significantly outperforming the previous SOTA by 32.5%.
****
Estimation and Quantization of Expected Persistence Diagrams
Vincent Divol, Theo Lacombe
Persistence diagrams (PDs) are the most common descriptors used to encode the topology of structured data appearing in challenging learning tasks; think e.g. of graphs, time series or point clouds sampled close to a manifold. Given random objects and the corresponding distribution of PDs, one may want to build a statistical summary—such as a mean—of these random PDs, which is however not a trivial task as the natural geometry of the space of PDs is not linear. In this article, we study two such summaries, the Expected Persistence Diagram (EPD), and its quantization. The EPD is a measure supported on $\mathbb{R}^2$, which may be approximated by its empirical counterpart. We prove that this estimator is optimal from a minimax standpoint on a large class of models with a parametric rate of convergence. The empirical EPD is simple and efficient to compute, but possibly has a very large support, hindering its use in practice. To overcome this issue, we propose an algorithm to compute a quantization of the empirical EPD, a measure with small support which is shown to approximate with near-optimal rates a quantization of the theoretical EPD.
****
On Energy-Based Models with Overparametrized Shallow Neural Networks
Carles Domingo-Enrich, Alberto Bietti, Eric Vanden-Eijnden, Joan Bruna
Energy-based models (EBMs) are a simple yet powerful framework for generative modeling. They are based on a trainable energy function which defines an associated Gibbs measure, and they can be trained and sampled from via well-established statistical tools, such as MCMC. Neural networks may be used as energy function approximators, providing both a rich class of expressive models as well as a flexible device to incorporate data structure. In this work we focus on shallow neural networks. Building from the incipient theory of overparametrized neural networks, we show that models trained in the so-called 'active' regime provide a statistical advantage over their associated 'lazy' or kernel regime, leading to improved adaptivity to hidden low-dimensional structure in the data distribution, as already observed in supervised learning. Our study covers both the maximum likelihood and Stein Discrepancy estimators, and we validate our theoretical results with numerical experiments on synthetic data.
****
Kernel-Based Reinforcement Learning: A Finite-Time Analysis
Omar Darwiche Domingues, Pierre Menard, Matteo Pirotta, Emilie Kaufmann, Michal Valko
We consider the exploration-exploitation dilemma in finite-horizon reinforcement learning problems whose state-action space is endowed with a metric. We introduce Kernel-UCBVI, a model-based optimistic algorithm that leverages the smoothness of the MDP and a non-parametric kernel estimator of the rewards and transitions to efficiently balance exploration and exploitation. For problems with K episodes and horizon H, we provide a regret bound of $\widetilde{O}\left(H^3 K^{\frac{2d}{2d+1}}\right)$, where d is the covering dimension of the joint state-action space. This is the first regret bound for kernel-based RL using smoothing kernels, which requires very weak assumptions on the MDP and applies to a wide range of tasks. We empirically validate our approach in continuous MDPs with sparse rewards.
****
Attention is not all you need: pure attention loses rank doubly exponentially with depth
Yihe Dong, Jean-Baptiste Cordonnier, Andreas Loukas
Attention-based architectures have become ubiquitous in machine learning. Yet, our understanding of the reasons for their effectiveness remains limited. This work proposes a new way to understand self-attention networks: we show that their output can be decomposed into a sum of smaller terms—or paths—each involving the

operation of a sequence of attention heads across layers. Using this path decomposition, we prove that self-attention possesses a strong inductive bias towards "token uniformity". Specifically, without skip connections or multi-layer perceptrons (MLPs), the output converges doubly exponentially to a rank-1 matrix. On the other hand, skip connections and MLPs stop the output from degeneration. Our experiments verify the convergence results on standard transformer architectures.
****
How rotational invariance of common kernels prevents generalization in high dimensions
Konstantin Donhauser, Mingqi Wu, Fanny Yang
Kernel ridge regression is well-known to achieve minimax optimal rates in low-dimensional settings. However, its behavior in high dimensions is much less understood. Recent work establishes consistency for high-dimensional kernel regression for a number of specific assumptions on the data distribution. In this paper, we show that in high dimensions, the rotational invariance property of commonly studied kernels (such as RBF, inner product kernels and fully-connected NTK of any depth) leads to inconsistent estimation unless the ground truth is a low-degree polynomial. Our lower bound on the generalization error holds for a wide range of distributions and kernels with different eigenvalue decays. This lower bound suggests that consistency results for kernel ridge regression in high dimensions generally require a more refined analysis that depends on the structure of the kernel beyond its eigenvalue decay.
****
Fast Stochastic Bregman Gradient Methods: Sharp Analysis and Variance Reduction
Radu Alexandru Dragomir, Mathieu Even, Hadrien Hendrikx
We study the problem of minimizing a relatively-smooth convex function using stochastic Bregman gradient methods. We first prove the convergence of Bregman Stochastic Gradient Descent (BSGD) to a region that depends on the noise (magnitude of the gradients) at the optimum. In particular, BSGD quickly converges to the exact minimizer when this noise is zero (interpolation setting, in which the data is fit perfectly). Otherwise, when the objective has a finite sum structure, we show that variance reduction can be used to counter the effect of noise. In particular, fast convergence to the exact minimizer can be obtained under additional regularity assumptions on the Bregman reference function. We illustrate the effectiveness of our approach on two key applications of relative smoothness: tomographic reconstruction with Poisson noise and statistical preconditioning for distributed optimization.
****
Bilinear Classes: A Structural Framework for Provable Generalization in RL
Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, Ruosong Wang
This work introduces Bilinear Classes, a new structural framework, which permit generalization in reinforcement learning in a wide variety of settings through the use of function approximation. The framework incorporates nearly all existing models in which a polynomial sample complexity is achievable, and, notably, also includes new models, such as the Linear Q*/V* model in which both the optimal Q-function and the optimal V-function are linear in some known feature space. Our main result provides an RL algorithm which has polynomial sample complexity for Bilinear Classes; notably, this sample complexity is stated in terms of a reduction to the generalization error of an underlying supervised learning sub-problem. These bounds nearly match the best known sample complexity bounds for existing models. Furthermore, this framework also extends to the infinite dimensional (RKHS) setting: for the the Linear Q*/V* model, linear MDPs, and linear mixture MDPs, we provide sample complexities that have no explicit dependence on the explicit feature dimension (which could be infinite), but instead depends only on information theoretic quantities.
****
Improved Contrastive Divergence Training of Energy-Based Models
Yilun Du, Shuang Li, Joshua Tenenbaum, Igor Mordatch

Contrastive divergence is a popular method of training energy-based models, but is known to have difficulties with training stability. We propose an adaptation to improve contrastive divergence training by scrutinizing a gradient term that is difficult to calculate and is often left out for convenience. We show that th is gradient term is numerically significant and in practice is important to avoi d training instabilities, while being tractable to estimate. We further highligh t how data augmentation and multi-scale processing can be used to improve model robustness and generation quality. Finally, we empirically evaluate stability of model architectures and show improved performance on a host of benchmarks and u se cases, such as image generation, OOD detection, and compositional generation.
****
Order-Agnostic Cross Entropy for Non-Autoregressive Machine Translation
Cunxiao Du, Zhaopeng Tu, Jing Jiang
We propose a new training objective named order-agnostic cross entropy (OaXE) fo r fully non-autoregressive translation (NAT) models. OaXE improves the standard cross-entropy loss to ameliorate the effect of word reordering, which is a commo n source of the critical multimodality problem in NAT. Concretely, OaXE removes the penalty for word order errors, and computes the cross entropy loss based on the best possible alignment between model predictions and target tokens. Since t he log loss is very sensitive to invalid references, we leverage cross entropy i nitialization and loss truncation to ensure the model focuses on a good part of the search space. Extensive experiments on major WMT benchmarks demonstrate that OaXE substantially improves translation performance, setting new state of the a rt for fully NAT models. Further analyses show that OaXE indeed alleviates the m ultimodality problem by reducing token repetitions and increasing prediction con fidence. Our code, data, and trained models are available at https://github.com/ tencent-ailab/ICML21_OAXE.
****
Putting the “Learning" into Learning-Augmented Algorithms for Frequency Estimati on
Elbert Du, Franklyn Wang, Michael Mitzenmacher
In learning-augmented algorithms, algorithms are enhanced using information from a machine learning algorithm. In turn, this suggests that we should tailor our machine-learning approach for the target algorithm. We here consider this synerg y in the context of the learned count-min sketch from (Hsu et al., 2019). Learni ng here is used to predict heavy hitters from a data stream, which are counted e xplicitly outside the sketch. We show that an approximately sufficient statistic for the performance of the underlying count-min sketch is given by the coverage of the predictor, or the normalized L^1 norm of keys that are filtered by the predictor to be explicitly counted. We show that machine learning models which are trained to optimize for coverage lead to large improvements in performance o ver prior approaches according to the average absolute frequency error. Our sour ce code can be found at https://github.com/franklynwang/putting-the-learning-in-LAA.
****
Estimating $α$-Rank from A Few Entries with Low Rank Matrix Completion
Yali Du, Xue Yan, Xu Chen, Jun Wang, Haifeng Zhang
Multi-agent evaluation aims at the assessment of an agent’s strategy on the basi s of interaction with others. Typically, existing methods such as α-rank and its approximation still require to exhaustively compare all pairs of joint s trategies for an accurate ranking, which in practice is computationally expensiv e. In this paper, we aim to reduce the number of pairwise comparisons in recover ing a satisfying ranking for n strategies in two-player meta-games, by explori ng the fact that agents with similar skills may achieve similar payoffs against others. Two situations are considered: the first one is when we can obtain the t rue payoffs; the other one is when we can only access noisy payoff. Based on the se formulations, we leverage low-rank matrix completion and design two novel alg orithms for noise-free and noisy evaluations respectively. For both of these set tings, we theorize that $O(nr \log n)$ (n is the number of agents and r is t he rank of the payoff matrix) payoff entries are required to achieve sufficientl

y well strategy evaluation performance. Empirical results on evaluating the strategies in three synthetic games and twelve real world games demonstrate that strategy evaluation from a few entries can lead to comparable performance to algorithms with full knowledge of the payoff matrix.
****

Learning Diverse-Structured Networks for Adversarial Robustness

Xuefeng Du, Jingfeng Zhang, Bo Han, Tongliang Liu, Yu Rong, Gang Niu, Junzhou Huang, Masashi Sugiyama

In adversarial training (AT), the main focus has been the objective and optimizer while the model has been less studied, so that the models being used are still those classic ones in standard training (ST). Classic network architectures (NAs) are generally worse than searched NA in ST, which should be the same in AT. In this paper, we argue that NA and AT cannot be handled independently, since given a dataset, the optimal NA in ST would be no longer optimal in AT. That being said, AT is time-consuming itself; if we directly search NAs in AT over large search spaces, the computation will be practically infeasible. Thus, we propose diverse-structured network (DS-Net), to significantly reduce the size of the search space: instead of low-level operations, we only consider predefined atomic blocks, where an atomic block is a time-tested building block like the residual block. There are only a few atomic blocks and thus we can weight all atomic blocks rather than find the best one in a searched block of DS-Net, which is an essential tradeoff between exploring diverse structures and exploiting the best structures. Empirical results demonstrate the advantages of DS-Net, i.e., weighting the atomic blocks.
****

Risk Bounds and Rademacher Complexity in Batch Reinforcement Learning

Yaqi Duan, Chi Jin, Zhiyuan Li

This paper considers batch Reinforcement Learning (RL) with general value function approximation. Our study investigates the minimal assumptions to reliably estimate/minimize Bellman error, and characterizes the generalization performance by (local) Rademacher complexities of general function classes, which makes initial steps in bridging the gap between statistical learning theory and batch RL. Concretely, we view the Bellman error as a surrogate loss for the optimality gap, and prove the followings: (1) In double sampling regime, the excess risk of Empirical Risk Minimizer (ERM) is bounded by the Rademacher complexity of the function class. (2) In the single sampling regime, sample-efficient risk minimization is not possible without further assumptions, regardless of algorithms. However, with completeness assumptions, the excess risk of FQI and a minimax style algorithm can be again bounded by the Rademacher complexity of the corresponding function classes. (3) Fast statistical rates can be achieved by using tools of local Rademacher complexity. Our analysis covers a wide range of function classes, including finite classes, linear spaces, kernel spaces, sparse linear features, etc.
****

Sawtooth Factorial Topic Embeddings Guided Gamma Belief Network

Zhibin Duan, Dongsheng Wang, Bo Chen, Chaojie Wang, Wenchao Chen, Yewen Li, Jie Ren, Mingyuan Zhou

Hierarchical topic models such as the gamma belief network (GBN) have delivered promising results in mining multi-layer document representations and discovering interpretable topic taxonomies. However, they often assume in the prior that the topics at each layer are independently drawn from the Dirichlet distribution, ignoring the dependencies between the topics both at the same layer and across different layers. To relax this assumption, we propose sawtooth factorial topic embedding guided GBN, a deep generative model of documents that captures the dependencies and semantic similarities between the topics in the embedding space. Specifically, both the words and topics are represented as embedding vectors of the same dimension. The topic matrix at a layer is factorized into the product of a factor loading matrix and a topic embedding matrix, the transpose of which is set as the factor loading matrix of the layer above. Repeating this particular type of factorization, which shares components between adjacent layers, leads to

a structure referred to as sawtooth factorization. An auto-encoding variational inference network is constructed to optimize the model parameter via stochastic gradient descent. Experiments on big corpora show that our models outperform other neural topic models on extracting deeper interpretable topics and deriving better document representations.
****
Exponential Reduction in Sample Complexity with Learning of Ising Model Dynamics
Arkopal Dutt, Andrey Lokhov, Marc D Vuffray, Sidhant Misra
The usual setting for learning the structure and parameters of a graphical model assumes the availability of independent samples produced from the corresponding multivariate probability distribution. However, for many models the mixing time of the respective Markov chain can be very large and i.i.d. samples may not be obtained. We study the problem of reconstructing binary graphical models from correlated samples produced by a dynamical process, which is natural in many applications. We analyze the sample complexity of two estimators that are based on the interaction screening objective and the conditional likelihood loss. We observe that for samples coming from a dynamical process far from equilibrium, the sample complexity reduces exponentially compared to a dynamical process that mixes quickly.
****
Reinforcement Learning Under Moral Uncertainty
Adrien Ecoffet, Joel Lehman
An ambitious goal for machine learning is to create agents that behave ethically: The capacity to abide by human moral norms would greatly expand the context in which autonomous agents could be practically and safely deployed, e.g. fully autonomous vehicles will encounter charged moral decisions that complicate their deployment. While ethical agents could be trained by rewarding correct behavior under a specific moral theory (e.g. utilitarianism), there remains widespread disagreement about the nature of morality. Acknowledging such disagreement, recent work in moral philosophy proposes that ethical behavior requires acting under moral uncertainty, i.e. to take into account when acting that one’s credence is split across several plausible ethical theories. This paper translates such insights to the field of reinforcement learning, proposes two training methods that realize different points among competing desiderata, and trains agents in simple environments to act under moral uncertainty. The results illustrate (1) how such uncertainty can help curb extreme behavior from commitment to single theories and (2) several technical complications arising from attempting to ground moral philosophy in RL (e.g. how can a principled trade-off between two competing but incomparable reward functions be reached). The aim is to catalyze progress towards morally-competent agents and highlight the potential of RL to contribute towards the computational grounding of moral philosophy.
****
Confidence-Budget Matching for Sequential Budgeted Learning
Yonathan Efroni, Nadav Merlis, Aadirupa Saha, Shie Mannor
A core element in decision-making under uncertainty is the feedback on the quality of the performed actions. However, in many applications, such feedback is restricted. For example, in recommendation systems, repeatedly asking the user to provide feedback on the quality of recommendations will annoy them. In this work, we formalize decision-making problems with querying budget, where there is a (possibly time-dependent) hard limit on the number of reward queries allowed. Specifically, we focus on multi-armed bandits, linear contextual bandits, and reinforcement learning problems. We start by analyzing the performance of ‘greedy’ algorithms that query a reward whenever they can. We show that in fully stochastic settings, doing so performs surprisingly well, but in the presence of any adversity, this might lead to linear regret. To overcome this issue, we propose the Confidence-Budget Matching (CBM) principle that queries rewards when the confidence intervals are wider than the inverse square root of the available budget. We analyze the performance of CBM based algorithms in different settings and show that it performs well in the presence of adversity in the contexts, initial states, and budgets.

****
Self-Paced Context Evaluation for Contextual Reinforcement Learning
Theresa Eimer, André Biedenkapp, Frank Hutter, Marius Lindauer
Reinforcement learning (RL) has made a lot of advances for solving a single problem in a given environment; but learning policies that generalize to unseen variations of a problem remains challenging. To improve sample efficiency for learning on such instances of a problem domain, we present Self-Paced Context Evaluation (SPaCE). Based on self-paced learning, SPaCE automatically generates instance curricula online with little computational overhead. To this end, SPaCE leverages information contained in state values during training to accelerate and improve training performance as well as generalization capabilities to new \tasks from the same problem domain. Nevertheless, SPaCE is independent of the problem domain at hand and can be applied on top of any RL agent with state-value function approximation. We demonstrate SPaCE's ability to speed up learning of different value-based RL agents on two environments, showing better generalization capabilities and up to 10x faster learning compared to naive approaches such as round robin or SPDRL, as the closest state-of-the-art approach.
****
Provably Strict Generalisation Benefit for Equivariant Models
Bryn Elesedy, Sheheryar Zaidi
It is widely believed that engineering a model to be invariant/equivariant improves generalisation. Despite the growing popularity of this approach, a precise characterisation of the generalisation benefit is lacking. By considering the simplest case of linear models, this paper provides the first provably non-zero improvement in generalisation for invariant/equivariant models when the target distribution is invariant/equivariant with respect to a compact group. Moreover, our work reveals an interesting relationship between generalisation, the number of training examples and properties of the group action. Our results rest on an observation of the structure of function spaces under averaging operators which, along with its consequences for feature averaging, may be of independent interest.
****
Efficient Iterative Amortized Inference for Learning Symmetric and Disentangled Multi-Object Representations
Patrick Emami, Pan He, Sanjay Ranka, Anand Rangarajan
Unsupervised multi-object representation learning depends on inductive biases to guide the discovery of object-centric representations that generalize. However, we observe that methods for learning these representations are either impractical due to long training times and large memory consumption or forego key inductive biases. In this work, we introduce EfficientMORL, an efficient framework for the unsupervised learning of object-centric representations. We show that optimization challenges caused by requiring both symmetry and disentanglement can in fact be addressed by high-cost iterative amortized inference by designing the framework to minimize its dependence on it. We take a two-stage approach to inference: first, a hierarchical variational autoencoder extracts symmetric and disentangled representations through bottom-up inference, and second, a lightweight network refines the representations with top-down feedback. The number of refinement steps taken during training is reduced following a curriculum, so that at test time with zero steps the model achieves 99.1% of the refined decomposition performance. We demonstrate strong object decomposition and disentanglement on the standard multi-object benchmark while achieving nearly an order of magnitude faster training and test time inference over the previous state-of-the-art model.
****
Implicit Bias of Linear RNNs
Melikasadat Emami, Mojtaba Sahraee-Ardakan, Parthe Pandit, Sundeep Rangan, Alyson K Fletcher
Contemporary wisdom based on empirical studies suggests that standard recurrent neural networks (RNNs) do not perform well on tasks requiring long-term memory. However, RNNs' poor ability to capture long-term dependencies has not been fully understood. This paper provides a rigorous explanation of this property in the special case of linear RNNs. Although this work is limited to linear RNNs, even

these systems have traditionally been difficult to analyze due to their non-linear parameterization. Using recently-developed kernel regime analysis, our main result shows that as the number of hidden units goes to infinity, linear RNNs learned from random initializations are functionally equivalent to a certain weighted 1D-convolutional network. Importantly, the weightings in the equivalent model cause an implicit bias to elements with smaller time lags in the convolution, and hence shorter memory. The degree of this bias depends on the variance of the transition matrix at initialization and is related to the classic exploding and vanishing gradients problem. The theory is validated with both synthetic and real data experiments.

****

Global Optimality Beyond Two Layers: Training Deep ReLU Networks via Convex Programs

Tolga Ergen, Mert Pilanci

Understanding the fundamental mechanism behind the success of deep neural networks is one of the key challenges in the modern machine learning literature. Despite numerous attempts, a solid theoretical analysis is yet to be developed. In this paper, we develop a novel unified framework to reveal a hidden regularization mechanism through the lens of convex optimization. We first show that the training of multiple three-layer ReLU sub-networks with weight decay regularization can be equivalently cast as a convex optimization problem in a higher dimensional space, where sparsity is enforced via a group ℓ_1-norm regularization. Consequently, ReLU networks can be interpreted as high dimensional feature selection methods. More importantly, we then prove that the equivalent convex problem can be globally optimized by a standard convex optimization solver with a polynomial-time complexity with respect to the number of samples and data dimension when the width of the network is fixed. Finally, we numerically validate our theoretical results via experiments involving both synthetic and real datasets.

****

Revealing the Structure of Deep Neural Networks via Convex Duality

Tolga Ergen, Mert Pilanci

We study regularized deep neural networks (DNNs) and introduce a convex analytic framework to characterize the structure of the hidden layers. We show that a set of optimal hidden layer weights for a norm regularized DNN training problem can be explicitly found as the extreme points of a convex set. For the special case of deep linear networks, we prove that each optimal weight matrix aligns with the previous layers via duality. More importantly, we apply the same characterization to deep ReLU networks with whitened data and prove the same weight alignment holds. As a corollary, we also prove that norm regularized deep ReLU networks yield spline interpolation for one-dimensional datasets which was previously known only for two-layer networks. Furthermore, we provide closed-form solutions for the optimal layer weights when data is rank-one or whitened. The same analysis also applies to architectures with batch normalization even for arbitrary data. Therefore, we obtain a complete explanation for a recent empirical observation termed Neural Collapse where class means collapse to the vertices of a simplex equiangular tight frame.

****

Whitening for Self-Supervised Representation Learning

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, Nicu Sebe

Most of the current self-supervised representation learning (SSL) methods are based on the contrastive loss and the instance-discrimination task, where augmented versions of the same image instance ("positives") are contrasted with instances extracted from other images ("negatives"). For the learning to be effective, many negatives should be compared with a positive pair, which is computationally demanding. In this paper, we propose a different direction and a new loss function for SSL, which is based on the whitening of the latent-space features. The whitening operation has a "scattering" effect on the batch samples, avoiding degenerate solutions where all the sample representations collapse to a single point. Our solution does not require asymmetric networks and it is conceptually simple. Moreover, since negatives are not needed, we can extract multiple positive pai

rs from the same image instance. The source code of the method and of all the experiments is available at: https://github.com/htdt/self-supervised.
****
Graph Mixture Density Networks
Federico Errica, Davide Bacciu, Alessio Micheli
We introduce the Graph Mixture Density Networks, a new family of machine learning models that can fit multimodal output distributions conditioned on graphs of arbitrary topology. By combining ideas from mixture models and graph representation learning, we address a broader class of challenging conditional density estimation problems that rely on structured data. In this respect, we evaluate our method on a new benchmark application that leverages random graphs for stochastic epidemic simulations. We show a significant improvement in the likelihood of epidemic outcomes when taking into account both multimodality and structure. The empirical analysis is complemented by two real-world regression tasks showing the effectiveness of our approach in modeling the output prediction uncertainty. Graph Mixture Density Networks open appealing research opportunities in the study of structure-dependent phenomena that exhibit non-trivial conditional output distributions.
****
Cross-Gradient Aggregation for Decentralized Learning from Non-IID Data
Yasaman Esfandiari, Sin Yong Tan, Zhanhong Jiang, Aditya Balu, Ethan Herron, Chinmay Hegde, Soumik Sarkar
Decentralized learning enables a group of collaborative agents to learn models using a distributed dataset without the need for a central parameter server. Recently, decentralized learning algorithms have demonstrated state-of-the-art results on benchmark data sets, comparable with centralized algorithms. However, the key assumption to achieve competitive performance is that the data is independently and identically distributed (IID) among the agents which, in real-life applications, is often not applicable. Inspired by ideas from continual learning, we propose Cross-Gradient Aggregation (CGA), a novel decentralized learning algorithm where (i) each agent aggregates cross-gradient information, i.e., derivatives of its model with respect to its neighbors' datasets, and (ii) updates its model using a projected gradient based on quadratic programming (QP). We theoretically analyze the convergence characteristics of CGA and demonstrate its efficiency on non-IID data distributions sampled from the MNIST and CIFAR-10 datasets. Our empirical comparisons show superior learning performance of CGA over existing state-of-the-art decentralized learning algorithms, as well as maintaining the improved performance under information compression to reduce peer-to-peer communication overhead. The code is available here on GitHub.
****
Weight-covariance alignment for adversarially robust neural networks
Panagiotis Eustratiadis, Henry Gouk, Da Li, Timothy Hospedales
Stochastic Neural Networks (SNNs) that inject noise into their hidden layers have recently been shown to achieve strong robustness against adversarial attacks. However, existing SNNs are usually heuristically motivated, and often rely on adversarial training, which is computationally costly. We propose a new SNN that achieves state-of-the-art performance without relying on adversarial training, and enjoys solid theoretical justification. Specifically, while existing SNNs inject learned or hand-tuned isotropic noise, our SNN learns an anisotropic noise distribution to optimize a learning-theoretic bound on adversarial robustness. We evaluate our method on a number of popular benchmarks, show that it can be applied to different architectures, and that it provides robustness to a variety of white-box and black-box attacks, while being simple and fast to train compared to existing alternatives.
****
Data augmentation for deep learning based accelerated MRI reconstruction with limited data
Zalan Fabian, Reinhard Heckel, Mahdi Soltanolkotabi
Deep neural networks have emerged as very successful tools for image restoration and reconstruction tasks. These networks are often trained end-to-end to direct

ly reconstruct an image from a noisy or corrupted measurement of that image. To achieve state-of-the-art performance, training on large and diverse sets of images is considered critical. However, it is often difficult and/or expensive to collect large amounts of training images. Inspired by the success of Data Augmentation (DA) for classification problems, in this paper, we propose a pipeline for data augmentation for accelerated MRI reconstruction and study its effectiveness at reducing the required training data in a variety of settings. Our DA pipeline, MRAugment, is specifically designed to utilize the invariances present in medical imaging measurements as naive DA strategies that neglect the physics of the problem fail. Through extensive studies on multiple datasets we demonstrate that in the low-data regime DA prevents overfitting and can match or even surpass the state of the art while using significantly fewer training data, whereas in the high-data regime it has diminishing returns. Furthermore, our findings show that DA improves the robustness of the model against various shifts in the test distribution.
****
Poisson-Randomised DirBN: Large Mutation is Needed in Dirichlet Belief Networks
Xuhui Fan, Bin Li, Yaqiong Li, Scott A. Sisson
The Dirichlet Belief Network (DirBN) was recently proposed as a promising deep generative model to learn interpretable deep latent distributions for objects. However, its current representation capability is limited since its latent distributions across different layers is prone to form similar patterns and can thus hardly use multi-layer structure to form flexible distributions. In this work, we propose Poisson-randomised Dirichlet Belief Networks (Pois-DirBN), which allows large mutations for the latent distributions across layers to enlarge the representation capability. Based on our key idea of inserting Poisson random variables in the layer-wise connection, Pois-DirBN first introduces a component-wise propagation mechanism to enable latent distributions to have large variations across different layers. Then, we develop a layer-wise Gibbs sampling algorithm to infer the latent distributions, leading to a larger number of effective layers compared to DirBN. In addition, we integrate out latent distributions and form a multi-stochastic deep integer network, which provides an alternative view on Pois-DirBN. We apply Pois-DirBN to relational modelling and validate its effectiveness through improved link prediction performance and more interpretable latent distribution visualisations. The code can be downloaded at https://github.com/xuhuifan/Pois_DirBN.
****
Model-based Reinforcement Learning for Continuous Control with Posterior Sampling
Ying Fan, Yifei Ming
Balancing exploration and exploitation is crucial in reinforcement learning (RL). In this paper, we study model-based posterior sampling for reinforcement learning (PSRL) in continuous state-action spaces theoretically and empirically. First, we show the first regret bound of PSRL in continuous spaces which is polynomial in the episode length to the best of our knowledge. With the assumption that reward and transition functions can be modeled by Bayesian linear regression, we develop a regret bound of $\tilde{O}(H^{3/2}d\sqrt{T})$, where H is the episode length, d is the dimension of the state-action space, and T indicates the total time steps. This result matches the best-known regret bound of non-PSRL methods in linear MDPs. Our bound can be extended to nonlinear cases as well with feature embedding: using linear kernels on the feature representation ϕ, the regret bound becomes $\tilde{O}(H^{3/2}d_{\phi}\sqrt{T})$, where d_ϕ is the dimension of the representation space. Moreover, we present MPC-PSRL, a model-based posterior sampling algorithm with model predictive control for action selection. To capture the uncertainty in models, we use Bayesian linear regression on the penultimate layer (the feature representation layer ϕ) of neural networks. Empirical results show that our algorithm achieves the state-of-the-art sample efficiency in benchmark continuous control tasks compared to prior model-based algorithms, and matches the asymptotic performance of model-free algorithms.

****
SECANT: Self-Expert Cloning for Zero-Shot Generalization of Visual Policies
Linxi Fan, Guanzhi Wang, De-An Huang, Zhiding Yu, Li Fei-Fei, Yuke Zhu, Animashr
ee Anandkumar
Generalization has been a long-standing challenge for reinforcement learning (RL
). Visual RL, in particular, can be easily distracted by irrelevant factors in h
igh-dimensional observation space. In this work, we consider robust policy learn
ing which targets zero-shot generalization to unseen visual environments with la
rge distributional shift. We propose SECANT, a novel self-expert cloning techniq
ue that leverages image augmentation in two stages to *decouple* robust represen
tation learning from policy optimization. Specifically, an expert policy is firs
t trained by RL from scratch with weak augmentations. A student network then lea
rns to mimic the expert policy by supervised learning with strong augmentations,
 making its representation more robust against visual variations compared to the
 expert. Extensive experiments demonstrate that SECANT significantly advances th
e state of the art in zero-shot generalization across 4 challenging domains. Our
 average reward improvements over prior SOTAs are: DeepMind Control (+26.5%), ro
botic manipulation (+337.8%), vision-based autonomous driving (+47.7%), and indo
or object navigation (+15.8%). Code release and video are available at https://l
inxifan.github.io/secant-site/.
****
On Estimation in Latent Variable Models
Guanhua Fang, Ping Li
Latent variable models have been playing a central role in statistics, economet
ics, machine learning with applications to repeated observation study, panel dat
a inference, user behavior analysis, etc. In many modern applications, the infer
ence based on latent variable models involves one or several of the following fe
atures: the presence of complex latent structure, the observed and latent variab
les being continuous or discrete, constraints on parameters, and data size being
 large. Therefore, solving an estimation problem for general latent variable mod
els is highly non-trivial. In this paper, we consider a gradient based method vi
a using variance reduction technique to accelerate estimation procedure. Theoret
ically, we show the convergence results for the proposed method under general an
d mild model assumptions. The algorithm has better computational complexity comp
ared with the classical gradient methods and maintains nice statistical properti
es. Various numerical results corroborate our theory.
****
On Variational Inference in Biclustering Models
Guanhua Fang, Ping Li
Biclustering structures exist ubiquitously in data matrices and the biclustering
 problem was first formalized by John Hartigan (1972) to cluster rows and column
s simultaneously. In this paper, we develop a theory for the estimation of gener
al biclustering models, where the data is assumed to follow certain statistical
distribution with underlying biclustering structure. Due to the existence of lat
ent variables, directly computing the maximal likelihood estimator is prohibitiv
ely difficult in practice and we instead consider the variational inference (VI)
 approach to solve the parameter estimation problem. Although variational infere
nce method generally has good empirical performance, there are very few theoreti
cal results around VI. In this paper, we obtain the precise estimation bound of
variational estimator and show that it matches the minimax rate in terms of esti
mation error under mild assumptions in biclustering setting. Furthermore, we stu
dy the convergence property of the coordinate ascent variational inference algor
ithm, where both local and global convergence results have been provided. Numeri
cal results validate our new theories.
****
Learning Bounds for Open-Set Learning
Zhen Fang, Jie Lu, Anjin Liu, Feng Liu, Guangquan Zhang
Traditional supervised learning aims to train a classifier in the closed-set wor
ld, where training and test samples share the same label space. In this paper, w
e target a more challenging and re_x0002_alistic setting: open-set learning (OSL

), where there exist test samples from the classes that are unseen during training. Although researchers have designed many methods from the algorith_x0002_mic perspectives, there are few methods that pro_x0002_vide generalization guarantees on their ability to achieve consistent performance on different train_x0002_ing samples drawn from the same distribution. Motivated by the transfer learning and probably approximate correct (PAC) theory, we make a bold attempt to study OSL by proving its general_x0002_ization error-given training samples with size n, the estimation error will get close to order Op(1/$\sqrt{}$n). This is the first study to provide a generalization bound for OSL, which we do by theoretically investigating the risk of the tar_x0002_get classifier on unknown classes. According to our theory, a novel algorithm, called auxiliary open-set risk (AOSR) is proposed to address the OSL problem. Experiments verify the efficacy of AOSR. The code is available at github.com/AnjinLiu/Openset_Learning_AOSR.
****
Streaming Bayesian Deep Tensor Factorization
Shikai Fang, Zheng Wang, Zhimeng Pan, Ji Liu, Shandian Zhe
Despite the success of existing tensor factorization methods, most of them conduct a multilinear decomposition, and rarely exploit powerful modeling frameworks, like deep neural networks, to capture a variety of complicated interactions in data. More important, for highly expressive, deep factorization, we lack an effective approach to handle streaming data, which are ubiquitous in real-world applications. To address these issues, we propose SBTD, a Streaming Bayesian Deep Tensor factorization method. We first use Bayesian neural networks (NNs) to build a deep tensor factorization model. We assign a spike-and-slab prior over each NN weight to encourage sparsity and to prevent overfitting. We then use multivariate Delta’s method and moment matching to approximate the posterior of the NN output and calculate the running model evidence, based on which we develop an efficient streaming posterior inference algorithm in the assumed-density-filtering and expectation propagation framework. Our algorithm provides responsive incremental updates for the posterior of the latent factors and NN weights upon receiving newly observed tensor entries, and meanwhile identify and inhibit redundant/useless weights. We show the advantages of our approach in four real-world applications.
****
PID Accelerated Value Iteration Algorithm
Amir-Massoud Farahmand, Mohammad Ghavamzadeh
The convergence rate of Value Iteration (VI), a fundamental procedure in dynamic programming and reinforcement learning, for solving MDPs can be slow when the discount factor is close to one. We propose modifications to VI in order to potentially accelerate its convergence behaviour. The key insight is the realization that the evolution of the value function approximations $(V_k)_{k \geq 0}$ in the VI procedure can be seen as a dynamical system. This opens up the possibility of using techniques from \emph{control theory} to modify, and potentially accelerate, this dynamics. We present such modifications based on simple controllers, such as PD (Proportional-Derivative), PI (Proportional-Integral), and PID. We present the error dynamics of these variants of VI, and provably (for certain classes of MDPs) and empirically (for more general classes) show that the convergence rate can be significantly improved. We also propose a gain adaptation mechanism in order to automatically select the controller gains, and empirically show the effectiveness of this procedure.
****
Near-Optimal Entrywise Anomaly Detection for Low-Rank Matrices with Sub-Exponential Noise
Vivek Farias, Andrew A Li, Tianyi Peng
We study the problem of identifying anomalies in a low-rank matrix observed with sub-exponential noise, motivated by applications in retail and inventory management. State of the art approaches to anomaly detection in low-rank matrices apparently fall short, since they require that non-anomalous entries be observed with vanishingly small noise (which is not the case in our problem, and indeed in many applications). So motivated, we propose a conceptually simple entrywise appr

oach to anomaly detection in low-rank matrices. Our approach accommodates a general class of probabilistic anomaly models. We extend recent work on entrywise error guarantees for matrix completion, establishing such guarantees for sub-exponential matrices, where in addition to missing entries, a fraction of entries are corrupted by (an also unknown) anomaly model. Viewing the anomaly detection as a classification task, to the best of our knowledge, we are the first to achieve the min-max optimal detection rate (up to log factors). Using data from a massive consumer goods retailer, we show that our approach provides significant improvements over incumbent approaches to anomaly detection.
****
Connecting Optimal Ex-Ante Collusion in Teams to Extensive-Form Correlation: Faster Algorithms and Positive Complexity Results
Gabriele Farina, Andrea Celli, Nicola Gatti, Tuomas Sandholm
We focus on the problem of finding an optimal strategy for a team of players that faces an opponent in an imperfect-information zero-sum extensive-form game. Team members are not allowed to communicate during play but can coordinate before the game. In this setting, it is known that the best the team can do is sample a profile of potentially randomized strategies (one per player) from a joint (a.k.a. correlated) probability distribution at the beginning of the game. In this paper, we first provide new modeling results about computing such an optimal distribution by drawing a connection to a different literature on extensive-form correlation. Second, we provide an algorithm that allows one for capping the number of profiles employed in the solution. This begets an anytime algorithm by increasing the cap. We find that often a handful of well-chosen such profiles suffices to reach optimal utility for the team. This enables team members to reach coordination through a simple and understandable plan. Finally, inspired by this observation and leveraging theoretical concepts that we introduce, we develop an efficient column-generation algorithm for finding an optimal distribution for the team. We evaluate it on a suite of common benchmark games. It is three orders of magnitude faster than the prior state of the art on games that the latter can solve and it can also solve several games that were previously unsolvable.
****
Train simultaneously, generalize better: Stability of gradient-based minimax learners
Farzan Farnia, Asuman Ozdaglar
The success of minimax learning problems of generative adversarial networks (GANs) has been observed to depend on the minimax optimization algorithm used for their training. This dependence is commonly attributed to the convergence speed and robustness properties of the underlying optimization algorithm. In this paper, we show that the optimization algorithm also plays a key role in the generalization performance of the trained minimax model. To this end, we analyze the generalization properties of standard gradient descent ascent (GDA) and proximal point method (PPM) algorithms through the lens of algorithmic stability as defined by Bousquet & Elisseeff, 2002 under both convex-concave and nonconvex-nonconcave minimax settings. While the GDA algorithm is not guaranteed to have a vanishing excess risk in convex-concave problems, we show the PPM algorithm enjoys a bounded excess risk in the same setup. For nonconvex-nonconcave problems, we compare the generalization performance of stochastic GDA and GDmax algorithms where the latter fully solves the maximization subproblem at every iteration. Our generalization analysis suggests the superiority of GDA provided that the minimization and maximization subproblems are solved simultaneously with similar learning rates. We discuss several numerical results indicating the role of optimization algorithms in the generalization of learned minimax models.
****
Unbalanced minibatch Optimal Transport; applications to Domain Adaptation
Kilian Fatras, Thibault Sejourne, Rémi Flamary, Nicolas Courty
Optimal transport distances have found many applications in machine learning for their capacity to compare non-parametric probability distributions. Yet their algorithmic complexity generally prevents their direct use on large scale datasets. Among the possible strategies to alleviate this issue, practitioners can rely

on computing estimates of these distances over subsets of data, i.e. minibatches. While computationally appealing, we highlight in this paper some limits of this strategy, arguing it can lead to undesirable smoothing effects. As an alternative, we suggest that the same minibatch strategy coupled with unbalanced optimal transport can yield more robust behaviors. We discuss the associated theoretical properties, such as unbiased estimators, existence of gradients and concentration bounds. Our experimental study shows that in challenging problems associated to domain adaptation, the use of unbalanced optimal transport leads to significantly better results, competing with or surpassing recent baselines.
****

Risk-Sensitive Reinforcement Learning with Function Approximation: A Debiasing Approach

Yingjie Fei, Zhuoran Yang, Zhaoran Wang

We study function approximation for episodic reinforcement learning with entropic risk measure. We first propose an algorithm with linear function approximation. Compared to existing algorithms, which suffer from improper regularization and regression biases, this algorithm features debiasing transformations in backward induction and regression procedures. We further propose an algorithm with general function approximation, which features implicit debiasing transformations. We prove that both algorithms achieve a sublinear regret and demonstrate a trade-off between generality and efficiency. Our analysis provides a unified framework for function approximation in risk-sensitive reinforcement learning, which leads to the first sublinear regret bounds in the setting.
****

Lossless Compression of Efficient Private Local Randomizers

Vitaly Feldman, Kunal Talwar

Locally Differentially Private (LDP) Reports are commonly used for collection of statistics and machine learning in the federated setting. In many cases the best known LDP algorithms require sending prohibitively large messages from the client device to the server (such as when constructing histograms over a large domain or learning a high-dimensional model). Here we demonstrate a general approach that, under standard cryptographic assumptions, compresses every efficient LDP algorithm with negligible loss in privacy and utility guarantees. The practical implication of our result is that in typical applications every message can be compressed to the size of the server’s pseudo-random generator seed. From this general approach we derive low-communication algorithms for the problems of frequency estimation and high-dimensional mean estimation. Our algorithms are simpler and more accurate than existing low-communication LDP algorithms for these well-studied problems.
****

Dimensionality Reduction for the Sum-of-Distances Metric

Zhili Feng, Praneeth Kacham, David Woodruff

We give a dimensionality reduction procedure to approximate the sum of distances of a given set of n points in R^d to any “shape” that lies in a k-dimensional subspace. Here, by “shape” we mean any set of points in R^d. Our algorithm takes an input in the form of an $n \times d$ matrix A, where each row of A denotes a data point, and outputs a subspace P of dimension $O(k^{3}/\epsilon^6)$ such that the projections of each of the n points onto the subspace P and the distances of each of the points to the subspace P are sufficient to obtain an ϵ-approximation to the sum of distances to any arbitrary shape that lies in a k-dimensional subspace of R^d. These include important problems such as k-median, k-subspace approximation, and (j,l) subspace clustering with $j \cdot l \leq k$. Dimensionality reduction reduces the data storage requirement to $(n+d)k^{3}/\epsilon^6$ from nnz(A). Here nnz(A) could potentially be as large as nd. Our algorithm runs in time nnz$(A)/\epsilon^2 + (n+d)$poly(k/ϵ), up to logarithmic factors. For dense matrices, where nnz$(A) \approx nd$, we give a faster algorithm, that runs in time $nd + (n+d)$poly(k/ϵ) up to logarithmic factors. Our dimensionality reduction algorithm can also be used to obtain poly(k/ϵ) size coresets for k-median and $(k,1)$-subspace approximation problems in polynomial time.

****
Reserve Price Optimization for First Price Auctions in Display Advertising
Zhe Feng, Sebastien Lahaie, Jon Schneider, Jinchao Ye
The display advertising industry has recently transitioned from second- to first -price auctions as its primary mechanism for ad allocation and pricing. In light of this, publishers need to re-evaluate and optimize their auction parameters, notably reserve prices. In this paper, we propose a gradient-based algorithm to adaptively update and optimize reserve prices based on estimates of bidders' responsiveness to experimental shocks in reserves. Our key innovation is to draw on the inherent structure of the revenue objective in order to reduce the variance of gradient estimates and improve convergence rates in both theory and practice. We show that revenue in a first-price auction can be usefully decomposed into a \emph{demand} component and a \emph{bidding} component, and introduce techniques to reduce the variance of each component. We characterize the bias-variance trade-offs of these techniques and validate the performance of our proposed algorithm through experiments on synthetic data and real display ad auctions data from a major ad exchange.
****
Uncertainty Principles of Encoding GANs
Ruili Feng, Zhouchen Lin, Jiapeng Zhu, Deli Zhao, Jingren Zhou, Zheng-Jun Zha
The compelling synthesis results of Generative Adversarial Networks (GANs) demonstrate rich semantic knowledge in their latent codes. To obtain this knowledge for downstream applications, encoding GANs has been proposed to learn encoders, such that real world data can be encoded to latent codes, which can be fed to generators to reconstruct those data. However, despite the theoretical guarantees of precise reconstruction in previous works, current algorithms generally reconstruct inputs with non-negligible deviations from inputs. In this paper we study this predicament of encoding GANs, which is indispensable research for the GAN community. We prove three uncertainty principles of encoding GANs in practice: a) the 'perfect' encoder and generator cannot be continuous at the same time, which implies that current framework of encoding GANs is ill-posed and needs rethinking; b) neural networks cannot approximate the underlying encoder and generator precisely at the same time, which explains why we cannot get 'perfect' encoders and generators as promised in previous theories; c) neural networks cannot be stable and accurate at the same time, which demonstrates the difficulty of training and trade-off between fidelity and disentanglement encountered in previous works. Our work may eliminate gaps between previous theories and empirical results, promote the understanding of GANs, and guide network designs for follow-up works.
****
Pointwise Binary Classification with Pairwise Confidence Comparisons
Lei Feng, Senlin Shu, Nan Lu, Bo Han, Miao Xu, Gang Niu, Bo An, Masashi Sugiyama
To alleviate the data requirement for training effective binary classifiers in binary classification, many weakly supervised learning settings have been proposed. Among them, some consider using pairwise but not pointwise labels, when pointwise labels are not accessible due to privacy, confidentiality, or security reasons. However, as a pairwise label denotes whether or not two data points share a pointwise label, it cannot be easily collected if either point is equally likely to be positive or negative. Thus, in this paper, we propose a novel setting called pairwise comparison (Pcomp) classification, where we have only pairs of unlabeled data that we know one is more likely to be positive than the other. Firstly, we give a Pcomp data generation process, derive an unbiased risk estimator (URE) with theoretical guarantee, and further improve URE using correction functions. Secondly, we link Pcomp classification to noisy-label learning to develop a progressive URE and improve it by imposing consistency regularization. Finally, we demonstrate by experiments the effectiveness of our methods, which suggests Pcomp is a valuable and practically useful type of pairwise supervision besides the pairwise label.
****
Provably Correct Optimization and Exploration with Non-linear Policies

Fei Feng, Wotao Yin, Alekh Agarwal, Lin Yang

Policy optimization methods remain a powerful workhorse in empirical Reinforcement Learning (RL), with a focus on neural policies that can easily reason over complex and continuous state and/or action spaces. Theoretical understanding of strategic exploration in policy-based methods with non-linear function approximation, however, is largely missing. In this paper, we address this question by designing ENIAC, an actor-critic method that allows non-linear function approximation in the critic. We show that under certain assumptions, e.g., a bounded eluder dimension d for the critic class, the learner finds to a near-optimal policy in $\widetilde{O}(\mathrm{poly}(d))$ exploration rounds. The method is robust to model misspecification and strictly extends existing works on linear function approximation. We also develop some computational optimizations of our approach with slightly worse statistical guarantees, and an empirical adaptation building on existing deep RL tools. We empirically evaluate this adaptation, and show that it outperforms prior heuristics inspired by linear methods, establishing the value in correctly reasoning about the agent’s uncertainty under non-linear function approximation.

****

KD3A: Unsupervised Multi-Source Decentralized Domain Adaptation via Knowledge Distillation

Haozhe Feng, Zhaoyang You, Minghao Chen, Tianye Zhang, Minfeng Zhu, Fei Wu, Chao Wu, Wei Chen

Conventional unsupervised multi-source domain adaptation (UMDA) methods assume all source domains can be accessed directly. However, this assumption neglects the privacy-preserving policy, where all the data and computations must be kept decentralized. There exist three challenges in this scenario: (1) Minimizing the domain distance requires the pairwise calculation of the data from the source and target domains, while the data on the source domain is not available. (2) The communication cost and privacy security limit the application of existing UMDA methods, such as the domain adversarial training. (3) Since users cannot govern the data quality, the irrelevant or malicious source domains are more likely to appear, which causes negative transfer. To address the above problems, we propose a privacy-preserving UMDA paradigm named Knowledge Distillation based Decentralized Domain Adaptation (KD3A), which performs domain adaptation through the knowledge distillation on models from different source domains. The extensive experiments show that KD3A significantly outperforms state-of-the-art UMDA approaches. Moreover, the KD3A is robust to the negative transfer and brings a 100x reduction of communication cost compared with other decentralized UMDA methods.

****

Understanding Noise Injection in GANs

Ruili Feng, Deli Zhao, Zheng-Jun Zha

Noise injection is an effective way of circumventing overfitting and enhancing generalization in machine learning, the rationale of which has been validated in deep learning as well. Recently, noise injection exhibits surprising effectiveness when generating high-fidelity images in Generative Adversarial Networks (GANs) (e.g. StyleGAN). Despite its successful applications in GANs, the mechanism of its validity is still unclear. In this paper, we propose a geometric framework to theoretically analyze the role of noise injection in GANs. First, we point out the existence of the adversarial dimension trap inherent in GANs, which leads to the difficulty of learning a proper generator. Second, we successfully model the noise injection framework with exponential maps based on Riemannian geometry. Guided by our theories, we propose a general geometric realization for noise injection. Under our novel framework, the simple noise injection used in StyleGAN reduces to the Euclidean case. The goal of our work is to make theoretical steps towards understanding the underlying mechanism of state-of-the-art GAN algorithms. Experiments on image generation and GAN inversion validate our theory in practice.

****

GNNAutoScale: Scalable and Expressive Graph Neural Networks via Historical Embeddings

Matthias Fey, Jan E. Lenssen, Frank Weichert, Jure Leskovec
We present GNNAutoScale (GAS), a framework for scaling arbitrary message-passing GNNs to large graphs. GAS prunes entire sub-trees of the computation graph by utilizing historical embeddings from prior training iterations, leading to constant GPU memory consumption in respect to input node size without dropping any data. While existing solutions weaken the expressive power of message passing due to sub-sampling of edges or non-trainable propagations, our approach is provably able to maintain the expressive power of the original GNN. We achieve this by providing approximation error bounds of historical embeddings and show how to tighten them in practice. Empirically, we show that the practical realization of our framework, PyGAS, an easy-to-use extension for PyTorch Geometric, is both fast and memory-efficient, learns expressive node representations, closely resembles the performance of their non-scaling counterparts, and reaches state-of-the-art performance on large-scale graphs.
****
PsiPhi-Learning: Reinforcement Learning with Demonstrations using Successor Features and Inverse Temporal Difference Learning
Angelos Filos, Clare Lyle, Yarin Gal, Sergey Levine, Natasha Jaques, Gregory Farquhar
We study reinforcement learning (RL) with no-reward demonstrations, a setting in which an RL agent has access to additional data from the interaction of other agents with the same environment. However, it has no access to the rewards or goals of these agents, and their objectives and levels of expertise may vary widely. These assumptions are common in multi-agent settings, such as autonomous driving. To effectively use this data, we turn to the framework of successor features. This allows us to disentangle shared features and dynamics of the environment from agent-specific rewards and policies. We propose a multi-task inverse reinforcement learning (IRL) algorithm, called \emph{inverse temporal difference learning} (ITD), that learns shared state features, alongside per-agent successor features and preference vectors, purely from demonstrations without reward labels. We further show how to seamlessly integrate ITD with learning from online environment interactions, arriving at a novel algorithm for reinforcement learning with demonstrations, called $\Psi \Phi$-learning (pronounced ‘Sci-Fi’). We provide empirical evidence for the effectiveness of $\Psi \Phi$-learning as a method for improving RL, IRL, imitation, and few-shot transfer, and derive worst-case bounds for its performance in zero-shot transfer to new tasks.
****
A Practical Method for Constructing Equivariant Multilayer Perceptrons for Arbitrary Matrix Groups
Marc Finzi, Max Welling, Andrew Gordon Wilson
Symmetries and equivariance are fundamental to the generalization of neural networks on domains such as images, graphs, and point clouds. Existing work has primarily focused on a small number of groups, such as the translation, rotation, and permutation groups. In this work we provide a completely general algorithm for solving for the equivariant layers of matrix groups. In addition to recovering solutions from other works as special cases, we construct multilayer perceptrons equivariant to multiple groups that have never been tackled before, including $\mathrm{O}(1,3)$, $\mathrm{O}(5)$, $\mathrm{Sp}(n)$, and the Rubik’s cube group. Our approach outperforms non-equivariant baselines, with applications to particle physics and modeling dynamical systems. We release our software library to enable researchers to construct equivariant layers for arbitrary
****
Few-Shot Conformal Prediction with Auxiliary Tasks
Adam Fisch, Tal Schuster, Tommi Jaakkola, Dr.Regina Barzilay
We develop a novel approach to conformal prediction when the target task has limited data available for training. Conformal prediction identifies a small set of promising output candidates in place of a single prediction, with guarantees that the set contains the correct answer with high probability. When training data is limited, however, the predicted set can easily become unusably large. In this work, we obtain substantially tighter prediction sets while maintaining desira

ble marginal guarantees by casting conformal prediction as a meta-learning paradigm over exchangeable collections of auxiliary tasks. Our conformalization algorithm is simple, fast, and agnostic to the choice of underlying model, learning algorithm, or dataset. We demonstrate the effectiveness of this approach across a number of few-shot classification and regression tasks in natural language processing, computer vision, and computational chemistry for drug discovery.
****
Scalable Certified Segmentation via Randomized Smoothing
Marc Fischer, Maximilian Baader, Martin Vechev
We present a new certification method for image and point cloud segmentation based on randomized smoothing. The method leverages a novel scalable algorithm for prediction and certification that correctly accounts for multiple testing, necessary for ensuring statistical guarantees. The key to our approach is reliance on established multiple-testing correction mechanisms as well as the ability to abstain from classifying single pixels or points while still robustly segmenting the overall input. Our experimental evaluation on synthetic data and challenging datasets, such as Pascal Context, Cityscapes, and ShapeNet, shows that our algorithm can achieve, for the first time, competitive accuracy and certification guarantees on real-world segmentation tasks. We provide an implementation at https://github.com/eth-sri/segmentation-smoothing.
****
What’s in the Box? Exploring the Inner Life of Neural Networks with Robust Rules
Jonas Fischer, Anna Olah, Jilles Vreeken
We propose a novel method for exploring how neurons within neural networks interact. In particular, we consider activation values of a network for given data, and propose to mine noise-robust rules of the form X {\rightarrow} Y , where X and Y are sets of neurons in different layers. We identify the best set of rules by the Minimum Description Length Principle as the rules that together are most descriptive of the activation data. To learn good rule sets in practice, we propose the unsupervised ExplaiNN algorithm. Extensive evaluation shows that the patterns it discovers give clear insight in how networks perceive the world: they identify shared, respectively class-specific traits, compositionality within the network, as well as locality in convolutional layers. Moreover, these patterns are not only easily interpretable, but also supercharge prototyping as they identify which groups of neurons to consider in unison.
****
Online Learning with Optimism and Delay
Genevieve E Flaspohler, Francesco Orabona, Judah Cohen, Soukayna Mouatadid, Miruna Oprescu, Paulo Orenstein, Lester Mackey
Inspired by the demands of real-time climate and weather forecasting, we develop optimistic online learning algorithms that require no parameter tuning and have optimal regret guarantees under delayed feedback. Our algorithms—DORM, DORM+, and AdaHedgeD—arise from a novel reduction of delayed online learning to optimistic online learning that reveals how optimistic hints can mitigate the regret penalty caused by delay. We pair this delay-as-optimism perspective with a new analysis of optimistic learning that exposes its robustness to hinting errors and a new meta-algorithm for learning effective hinting strategies in the presence of delay. We conclude by benchmarking our algorithms on four subseasonal climate forecasting tasks, demonstrating low regret relative to state-of-the-art forecasting models.
****
Online A-Optimal Design and Active Linear Regression
Xavier Fontaine, Pierre Perrault, Michal Valko, Vianney Perchet
We consider in this paper the problem of optimal experiment design where a decision maker can choose which points to sample to obtain an estimate $\hat{\beta}$ of the hidden parameter $\beta^{\star}$ of an underlying linear model. The key challenge of this work lies in the heteroscedasticity assumption that we make, meaning that each covariate has a different and unknown variance. The goal of the decision maker is then to figure out on the fly the optimal way to allocate the total budget of T samples between covariates, as sampling several times a spec

ific one will reduce the variance of the estimated model around it (but at the cost of a possible higher variance elsewhere). By trying to minimize the ℓ^2-loss $\mathbb{E} [\lVert\hat{\beta}-\beta^{\star}\rVert^2]$ the decision maker is actually minimizing the trace of the covariance matrix of the problem, which corresponds then to online A-optimal design. Combining techniques from bandit and convex optimization we propose a new active sampling algorithm and we compare it with existing ones. We provide theoretical guarantees of this algorithm in different settings, including a $\mathcal{O}(T^{-2})$ regret bound in the case where the covariates form a basis of the feature space, generalizing and improving existing results. Numerical experiments validate our theoretical findings.
****
Deep Adaptive Design: Amortizing Sequential Bayesian Experimental Design
Adam Foster, Desi R Ivanova, Ilyas Malik, Tom Rainforth
We introduce Deep Adaptive Design (DAD), a method for amortizing the cost of adaptive Bayesian experimental design that allows experiments to be run in real-time. Traditional sequential Bayesian optimal experimental design approaches require substantial computation at each stage of the experiment. This makes them unsuitable for most real-world applications, where decisions must typically be made quickly. DAD addresses this restriction by learning an amortized design network upfront and then using this to rapidly run (multiple) adaptive experiments at deployment time. This network represents a design policy which takes as input the data from previous steps, and outputs the next design using a single forward pass; these design decisions can be made in milliseconds during the live experiment. To train the network, we introduce contrastive information bounds that are suitable objectives for the sequential setting, and propose a customized network architecture that exploits key symmetries. We demonstrate that DAD successfully amortizes the process of experimental design, outperforming alternative strategies on a number of problems.
****
Efficient Online Learning for Dynamic k-Clustering
Dimitris Fotakis, Georgios Piliouras, Stratis Skoulakis
In this work, we study dynamic clustering problems from the perspective of online learning. We consider an online learning problem, called \textit{Dynamic k-Clustering}, in which k centers are maintained in a metric space over time (centers may change positions) such as a dynamically changing set of r clients is served in the best possible way. The connection cost at round t is given by the \textit{p-norm} of the vector formed by the distance of each client to its closest center at round t, for some $p\geq 1$. We design a \textit{$\Theta\left(\min(k,r) \right)$-regret} polynomial-time online learning algorithm, while we show that, under some well-established computational complexity conjectures, \textit{constant-regret} cannot be achieved in polynomial-time. In addition to the efficient solution of Dynamic k-Clustering, our work contributes to the long line of research of combinatorial online learning.
****
Clustered Sampling: Low-Variance and Improved Representativity for Clients Selection in Federated Learning
Yann Fraboni, Richard Vidal, Laetitia Kameni, Marco Lorenzi
This work addresses the problem of optimizing communications between server and clients in federated learning (FL). Current sampling approaches in FL are either biased, or non optimal in terms of server-clients communications and training stability. To overcome this issue, we introduce clustered sampling for clients selection. We prove that clustered sampling leads to better clients representativity and to reduced variance of the clients stochastic aggregation weights in FL. Compatibly with our theory, we provide two different clustering approaches enabling clients aggregation based on 1) sample size, and 2) models similarity. Through a series of experiments in non-iid and unbalanced scenarios, we demonstrate that model aggregation through clustered sampling consistently leads to better training convergence and variability when compared to standard sampling approaches. Our approach does not require any additional operation on the clients side, and can be seamlessly integrated in standard FL implementations. Finally, cluste

red sampling is compatible with existing methods and technologies for privacy en hancement, and for communication reduction through model compression.
****
Agnostic Learning of Halfspaces with Gradient Descent via Soft Margins
Spencer Frei, Yuan Cao, Quanquan Gu
We analyze the properties of gradient descent on convex surrogates for the zero-one loss for the agnostic learning of halfspaces. We show that when a quantity w e refer to as the \textit{soft margin} is well-behaved—a condition satisfied by log-concave isotropic distributions among others—minimizers of convex surrogates for the zero-one loss are approximate minimizers for the zero-one loss itself. As standard convex optimization arguments lead to efficient guarantees for minim izing convex surrogates of the zero-one loss, our methods allow for the first po sitive guarantees for the classification error of halfspaces learned by gradient descent using the binary cross-entropy or hinge loss in the presence of agnosti c label noise.
****
Provable Generalization of SGD-trained Neural Networks of Any Width in the Prese nce of Adversarial Label Noise
Spencer Frei, Yuan Cao, Quanquan Gu
We consider a one-hidden-layer leaky ReLU network of arbitrary width trained by stochastic gradient descent (SGD) following an arbitrary initialization. We prov e that SGD produces neural networks that have classification accuracy competitiv e with that of the best halfspace over the distribution for a broad class of dis tributions that includes log-concave isotropic and hard margin distributions. Eq uivalently, such networks can generalize when the data distribution is linearly separable but corrupted with adversarial label noise, despite the capacity to ov erfit. To the best of our knowledge, this is the first work to show that overpar ameterized neural networks trained by SGD can generalize when the data is corrup ted with adversarial label noise.
****
Post-selection inference with HSIC-Lasso
Tobias Freidling, Benjamin Poignard, Héctor Climente-González, Makoto Yamada
Detecting influential features in non-linear and/or high-dimensional data is a c hallenging and increasingly important task in machine learning. Variable selecti on methods have thus been gaining much attention as well as post-selection infer ence. Indeed, the selected features can be significantly flawed when the selecti on procedure is not accounted for. We propose a selective inference procedure us ing the so-called model-free "HSIC-Lasso" based on the framework of truncated Ga ussians combined with the polyhedral lemma. We then develop an algorithm, which allows for low computational costs and provides a selection of the regularisatio n parameter. The performance of our method is illustrated by both artificial and real-world data based experiments, which emphasise a tight control of the type-I error, even for small sample sizes.
****
Variational Data Assimilation with a Learned Inverse Observation Operator
Thomas Frerix, Dmitrii Kochkov, Jamie Smith, Daniel Cremers, Michael Brenner, St ephan Hoyer
Variational data assimilation optimizes for an initial state of a dynamical syst em such that its evolution fits observational data. The physical model can subse quently be evolved into the future to make predictions. This principle is a corn erstone of large scale forecasting applications such as numerical weather predic tion. As such, it is implemented in current operational systems of weather forec asting agencies across the globe. However, finding a good initial state poses a difficult optimization problem in part due to the non-invertible relationship be tween physical states and their corresponding observations. We learn a mapping f rom observational data to physical states and show how it can be used to improve optimizability. We employ this mapping in two ways: to better initialize the no n-convex optimization problem, and to reformulate the objective function in bett er behaved physics space instead of observation space. Our experimental results for the Lorenz96 model and a two-dimensional turbulent fluid flow demonstrate th

at this procedure significantly improves forecast quality for chaotic systems.
****
Bayesian Quadrature on Riemannian Data Manifolds
Christian Fröhlich, Alexandra Gessner, Philipp Hennig, Bernhard Schölkopf, Georgios Arvanitidis
Riemannian manifolds provide a principled way to model nonlinear geometric structure inherent in data. A Riemannian metric on said manifolds determines geometry-aware shortest paths and provides the means to define statistical models accordingly. However, these operations are typically computationally demanding. To ease this computational burden, we advocate probabilistic numerical methods for Riemannian statistics. In particular, we focus on Bayesian quadrature (BQ) to numerically compute integrals over normal laws on Riemannian manifolds learned from data. In this task, each function evaluation relies on the solution of an expensive initial value problem. We show that by leveraging both prior knowledge and an active exploration scheme, BQ significantly reduces the number of required evaluations and thus outperforms Monte Carlo methods on a wide range of integration problems. As a concrete application, we highlight the merits of adopting Riemannian geometry with our proposed framework on a nonlinear dataset from molecular dynamics.
****
Learn-to-Share: A Hardware-friendly Transfer Learning Framework Exploiting Computation and Parameter Sharing
Cheng Fu, Hanxian Huang, Xinyun Chen, Yuandong Tian, Jishen Zhao
Task-specific fine-tuning on pre-trained transformers has achieved performance breakthroughs in multiple NLP tasks. Yet, as both computation and parameter size grows linearly with the number of sub-tasks, it is increasingly difficult to adopt such methods to the real world due to unrealistic memory and computation overhead on computing devices. Previous works on fine-tuning focus on reducing the growing parameter size to save storage cost by parameter sharing. However, compared to storage, the constraint of computation is a more critical issue with the fine-tuning models in modern computing environments. In this work, we propose LeTS, a framework that leverages both computation and parameter sharing across multiple tasks. Compared to traditional fine-tuning, LeTS proposes a novel neural architecture that contains a fixed pre-trained transformer model, plus learnable additive components for sub-tasks. The learnable components reuse the intermediate activations in the fixed pre-trained model, decoupling computation dependency. Differentiable neural architecture search is used to determine a task-specific computation sharing scheme, and a novel early stage pruning is applied to additive components for sparsity to achieve parameter sharing. Extensive experiments show that with 1.4% of extra parameters per task, LeTS reduces the computation by 49.5% on GLUE benchmarks with only 0.2% accuracy loss compared to full fine-tuning.
****
Learning Task Informed Abstractions
Xiang Fu, Ge Yang, Pulkit Agrawal, Tommi Jaakkola
Current model-based reinforcement learning methods struggle when operating from complex visual scenes due to their inability to prioritize task-relevant features. To mitigate this problem, we propose learning Task Informed Abstractions (TIA) that explicitly separates reward-correlated visual features from distractors. For learning TIA, we introduce the formalism of Task Informed MDP (TiMDP) that is realized by training two models that learn visual features via cooperative reconstruction, but one model is adversarially dissociated from the reward signal. Empirical evaluation shows that TIA leads to significant performance gains over state-of-the-art methods on many visual control tasks where natural and unconstrained visual distractions pose a formidable challenge. Project page: https://xiangfu.co/tia
****
Double-Win Quant: Aggressively Winning Robustness of Quantized Deep Neural Networks via Random Precision Training and Inference
Yonggan Fu, Qixuan Yu, Meng Li, Vikas Chandra, Yingyan Lin

Quantization is promising in enabling powerful yet complex deep neural networks (DNNs) to be deployed into resource constrained platforms. However, quantized DNNs are vulnerable to adversarial attacks unless being equipped with sophisticated techniques, leading to a dilemma of struggling between DNNs’ efficiency and robustness. In this work, we demonstrate a new perspective regarding quantization’s role in DNNs’ robustness, advocating that quantization can be leveraged to largely boost DNNs’ robustness, and propose a framework dubbed Double-Win Quant that can boost the robustness of quantized DNNs over their full precision counterparts by a large margin. Specifically, we for the first time identify that when an adversarially trained model is quantized to different precisions in a post-training manner, the associated adversarial attacks transfer poorly between different precisions. Leveraging this intriguing observation, we further develop Double-Win Quant integrating random precision inference and training to further reduce and utilize the poor adversarial transferability, enabling an aggressive “win-win" in terms of DNNs’ robustness and efficiency. Extensive experiments and ablation studies consistently validate Double-Win Quant’s effectiveness and advantages over state-of-the-art (SOTA) adversarial training methods across various attacks/models/datasets. Our codes are available at: https://github.com/RICE-EIC/Double-Win-Quant.
****
Auto-NBA: Efficient and Effective Search Over the Joint Space of Networks, Bitwidths, and Accelerators
Yonggan Fu, Yongan Zhang, Yang Zhang, David Cox, Yingyan Lin
While maximizing deep neural networks’ (DNNs’) acceleration efficiency requires a joint search/design of three different yet highly coupled aspects, including the networks, bitwidths, and accelerators, the challenges associated with such a joint search have not yet been fully understood and addressed. The key challenges include (1) the dilemma of whether to explode the memory consumption due to the huge joint space or achieve sub-optimal designs, (2) the discrete nature of the accelerator design space that is coupled yet different from that of the networks and bitwidths, and (3) the chicken and egg problem associated with network-accelerator co-search, i.e., co-search requires operation-wise hardware cost, which is lacking during search as the optimal accelerator depending on the whole network is still unknown during search. To tackle these daunting challenges towards optimal and fast development of DNN accelerators, we propose a framework dubbed Auto-NBA to enable jointly searching for the Networks, Bitwidths, and Accelerators, by efficiently localizing the optimal design within the huge joint design space for each target dataset and acceleration specification. Our Auto-NBA integrates a heterogeneous sampling strategy to achieve unbiased search with constant memory consumption, and a novel joint-search pipeline equipped with a generic differentiable accelerator search engine. Extensive experiments and ablation studies validate that both Auto-NBA generated networks and accelerators consistently outperform state-of-the-art designs (including co-search/exploration techniques, hardware-aware NAS methods, and DNN accelerators), in terms of search time, task accuracy, and accelerator efficiency. Our codes are available at: https://github.com/RICE-EIC/Auto-NBA.
****
A Deep Reinforcement Learning Approach to Marginalized Importance Sampling with the Successor Representation
Scott Fujimoto, David Meger, Doina Precup
Marginalized importance sampling (MIS), which measures the density ratio between the state-action occupancy of a target policy and that of a sampling distribution, is a promising approach for off-policy evaluation. However, current state-of-the-art MIS methods rely on complex optimization tricks and succeed mostly on simple toy problems. We bridge the gap between MIS and deep reinforcement learning by observing that the density ratio can be computed from the successor representation of the target policy. The successor representation can be trained through deep reinforcement learning methodology and decouples the reward optimization from the dynamics of the environment, making the resulting algorithm stable and applicable to high-dimensional domains. We evaluate the empirical performance of

our approach on a variety of challenging Atari and MuJoCo environments.
****

Learning disentangled representations via product manifold projection
Marco Fumero, Luca Cosmo, Simone Melzi, Emanuele Rodola
We propose a novel approach to disentangle the generative factors of variation underlying a given set of observations. Our method builds upon the idea that the (unknown) low-dimensional manifold underlying the data space can be explicitly modeled as a product of submanifolds. This definition of disentanglement gives rise to a novel weakly-supervised algorithm for recovering the unknown explanatory factors behind the data. At training time, our algorithm only requires pairs of non i.i.d. data samples whose elements share at least one, possibly multidimensional, generative factor of variation. We require no knowledge on the nature of these transformations, and do not make any limiting assumption on the properties of each subspace. Our approach is easy to implement, and can be successfully applied to different kinds of data (from images to 3D surfaces) undergoing arbitrary transformations. In addition to standard synthetic benchmarks, we showcase our method in challenging real-world applications, where we compare favorably with the state of the art.
****

Policy Information Capacity: Information-Theoretic Measure for Task Complexity in Deep Reinforcement Learning
Hiroki Furuta, Tatsuya Matsushima, Tadashi Kozuno, Yutaka Matsuo, Sergey Levine, Ofir Nachum, Shixiang Shane Gu
Progress in deep reinforcement learning (RL) research is largely enabled by benchmark task environments. However, analyzing the nature of those environments is often overlooked. In particular, we still do not have agreeable ways to measure the difficulty or solvability of a task, given that each has fundamentally different actions, observations, dynamics, rewards, and can be tackled with diverse RL algorithms. In this work, we propose policy information capacity (PIC) – the mutual information between policy parameters and episodic return – and policy-optimal information capacity (POIC) – between policy parameters and episodic optimality – as two environment-agnostic, algorithm-agnostic quantitative metrics for task difficulty. Evaluating our metrics across toy environments as well as continuous control benchmark tasks from OpenAI Gym and DeepMind Control Suite, we empirically demonstrate that these information-theoretic metrics have higher correlations with normalized task solvability scores than a variety of alternatives. Lastly, we show that these metrics can also be used for fast and compute-efficient optimizations of key design parameters such as reward shaping, policy architectures, and MDP properties for better solvability by RL algorithms without ever running full RL experiments.
****

An Information-Geometric Distance on the Space of Tasks
Yansong Gao, Pratik Chaudhari
This paper prescribes a distance between learning tasks modeled as joint distributions on data and labels. Using tools in information geometry, the distance is defined to be the length of the shortest weight trajectory on a Riemannian manifold as a classifier is fitted on an interpolated task. The interpolated task evolves from the source to the target task using an optimal transport formulation. This distance, which we call the "coupled transfer distance" can be compared across different classifier architectures. We develop an algorithm to compute the distance which iteratively transports the marginal on the data of the source task to that of the target task while updating the weights of the classifier to track this evolving data distribution. We develop theory to show that our distance captures the intuitive idea that a good transfer trajectory is the one that keeps the generalization gap small during transfer, in particular at the end on the target task. We perform thorough empirical validation and analysis across diverse image classification datasets to show that the coupled transfer distance correlates strongly with the difficulty of fine-tuning.
****

Maximum Mean Discrepancy Test is Aware of Adversarial Attacks

Ruize Gao, Feng Liu, Jingfeng Zhang, Bo Han, Tongliang Liu, Gang Niu, Masashi Sugiyama
The maximum mean discrepancy (MMD) test could in principle detect any distributional discrepancy between two datasets. However, it has been shown that the MMD test is unaware of adversarial attacks–the MMD test failed to detect the discrepancy between natural data and adversarial data. Given this phenomenon, we raise a question: are natural and adversarial data really from different distributions? The answer is affirmative–the previous use of the MMD test on the purpose missed three key factors, and accordingly, we propose three components. Firstly, the Gaussian kernel has limited representation power, and we replace it with an effective deep kernel. Secondly, the test power of the MMD test was neglected, and we maximize it following asymptotic statistics. Finally, adversarial data may be non-independent, and we overcome this issue with the help of wild bootstrap. By taking care of the three factors, we verify that the MMD test is aware of adversarial attacks, which lights up a novel road for adversarial data detection based on two-sample tests.
****
Unsupervised Co-part Segmentation through Assembly
Qingzhe Gao, Bin Wang, Libin Liu, Baoquan Chen
Co-part segmentation is an important problem in computer vision for its rich applications. We propose an unsupervised learning approach for co-part segmentation from images. For the training stage, we leverage motion information embedded in videos and explicitly extract latent representations to segment meaningful object parts. More importantly, we introduce a dual procedure of part-assembly to form a closed loop with part-segmentation, enabling an effective self-supervision. We demonstrate the effectiveness of our approach with a host of extensive experiments, ranging from human bodies, hands, quadruped, and robot arms. We show that our approach can achieve meaningful and compact part segmentation, outperforming state-of-the-art approaches on diverse benchmarks.
****
Discriminative Complementary-Label Learning with Weighted Loss
Yi Gao, Min-Ling Zhang
Complementary-label learning (CLL) deals with the weak supervision scenario where each training instance is associated with one \emph{complementary} label, which specifies the class label that the instance does \emph{not} belong to. Given the training instance ${\bm x}$, existing CLL approaches aim at modeling the \emph{generative} relationship between the complementary label $\bar y$, i.e. $P(\bar y\mid {\bm x})$, and the ground-truth label y, i.e. $P(y\mid {\bm x})$. Nonetheless, as the ground-truth label is not directly accessible for complementarily labeled training instance, strong generative assumptions may not hold for real-world CLL tasks. In this paper, we derive a simple and theoretically-sound \emph{discriminative} model towards $P(\bar y\mid {\bm x})$, which naturally leads to a risk estimator with estimation error bound at $\mathcal{O}(1/\sqrt{n})$ convergence rate. Accordingly, a practical CLL approach is proposed by further introducing weighted loss to the empirical risk to maximize the predictive gap between potential ground-truth label and complementary label. Extensive experiments clearly validate the effectiveness of the proposed discriminative complementary-label learning approach.
****
RATT: Leveraging Unlabeled Data to Guarantee Generalization
Saurabh Garg, Sivaraman Balakrishnan, Zico Kolter, Zachary Lipton
To assess generalization, machine learning scientists typically either (i) bound the generalization gap and then (after training) plug in the empirical risk to obtain a bound on the true risk; or (ii) validate empirically on holdout data. However, (i) typically yields vacuous guarantees for overparameterized models; and (ii) shrinks the training set and its guarantee erodes with each re-use of the holdout set. In this paper, we leverage unlabeled data to produce generalization bounds. After augmenting our (labeled) training set with randomly labeled data, we train in the standard fashion. Whenever classifiers achieve low error on the clean data but high error on the random data, our bound ensures that the true

risk is low. We prove that our bound is valid for 0-1 empirical risk minimization and with linear classifiers trained by gradient descent. Our approach is especially useful in conjunction with deep learning due to the early learning phenomenon whereby networks fit true labels before noisy labels but requires one intuitive assumption. Empirically, on canonical computer vision and NLP tasks, our bound provides non-vacuous generalization guarantees that track actual performance closely. This work enables practitioners to certify generalization even when (labeled) holdout data is unavailable and provides insights into the relationship between random label noise and generalization.
****
On Proximal Policy Optimization’s Heavy-tailed Gradients
Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, Zico Kolter, Zachary Lipton, Sivaraman Balakrishnan, Ruslan Salakhutdinov, Pradeep Ravikumar
Modern policy gradient algorithms such as Proximal Policy Optimization (PPO) rely on an arsenal of heuristics, including loss clipping and gradient clipping, to ensure successful learning. These heuristics are reminiscent of techniques from robust statistics, commonly used for estimation in outlier-rich ("heavy-tailed") regimes. In this paper, we present a detailed empirical study to characterize the heavy-tailed nature of the gradients of the PPO surrogate reward function. We demonstrate that the gradients, especially for the actor network, exhibit pronounced heavy-tailedness and that it increases as the agent’s policy diverges from the behavioral policy (i.e., as the agent goes further off policy). Further examination implicates the likelihood ratios and advantages in the surrogate reward as the main sources of the observed heavy-tailedness. We then highlight issues arising due to the heavy-tailed nature of the gradients. In this light, we study the effects of the standard PPO clipping heuristics, demonstrating that these tricks primarily serve to offset heavy-tailedness in gradients. Thus motivated, we propose incorporating GMOM, a high-dimensional robust estimator, into PPO as a substitute for three clipping tricks. Despite requiring less hyperparameter tuning, our method matches the performance of PPO (with all heuristics enabled) on a battery of MuJoCo continuous control tasks.
****
What does LIME really see in images?
Damien Garreau, Dina Mardaoui
The performance of modern algorithms on certain computer vision tasks such as object recognition is now close to that of humans. This success was achieved at the price of complicated architectures depending on millions of parameters and it has become quite challenging to understand how particular predictions are made. Interpretability methods propose to give us this understanding. In this paper, we study LIME, perhaps one of the most popular. On the theoretical side, we show that when the number of generated examples is large, LIME explanations are concentrated around a limit explanation for which we give an explicit expression. We further this study for elementary shape detectors and linear models. As a consequence of this analysis, we uncover a connection between LIME and integrated gradients, another explanation method. More precisely, the LIME explanations are similar to the sum of integrated gradients over the superpixels used in the preprocessing step of LIME.
****
Parametric Graph for Unimodal Ranking Bandit
Camille-Sovanneary Gauthier, Romaric Gaudel, Elisa Fromont, Boammani Aser Lompo
We tackle the online ranking problem of assigning L items to K positions on a web page in order to maximize the number of user clicks. We propose an original algorithm, easy to implement and with strong theoretical guarantees to tackle this problem in the Position-Based Model (PBM) setting, well suited for applications where items are displayed on a grid. Besides learning to rank, our algorithm, GRAB (for parametric Graph for unimodal RAnking Bandit), also learns the parameter of a compact graph over permutations of K items among L. The logarithmic regret bound of this algorithm is a direct consequence of the unimodality property of the bandit setting with respect to the learned graph. Experiments against state-of-the-art learning algorithms which also tackle the PBM setting, show

that our method is more efficient while giving regret performance on par with the best known algorithms on simulated and real life datasets.
****
Let’s Agree to Degree: Comparing Graph Convolutional Networks in the Message-Passing Framework
Floris Geerts, Filip Mazowiecki, Guillermo Perez
In this paper we cast neural networks defined on graphs as message-passing neural networks (MPNNs) to study the distinguishing power of different classes of such models. We are interested in when certain architectures are able to tell vertices apart based on the feature labels given as input with the graph. We consider two variants of MPNNS: anonymous MPNNs whose message functions depend only on the labels of vertices involved; and degree-aware MPNNs whose message functions can additionally use information regarding the degree of vertices. The former class covers popular graph neural network (GNN) formalisms for which the distinguished power is known. The latter covers graph convolutional networks (GCNs), introduced by Kipf and Welling, for which the distinguishing power was unknown. We obtain lower and upper bounds on the distinguishing power of (anonymous and degree-aware) MPNNs in terms of the distinguishing power of the Weisfeiler-Lehman (WL) algorithm. Our main results imply that (i) the distinguishing power of GCNs is bounded by the WL algorithm, but they may be one step ahead; (ii) the WL algorithm cannot be simulated by “plain vanilla” GCNs but the addition of a trade-off parameter between features of the vertex and those of its neighbours (as proposed by Kipf and Welling) resolves this problem.
****
On the difficulty of unbiased alpha divergence minimization
Tomas Geffner, Justin Domke
Several approximate inference algorithms have been proposed to minimize an alpha-divergence between an approximating distribution and a target distribution. Many of these algorithms introduce bias, the magnitude of which becomes problematic in high dimensions. Other algorithms are unbiased. These often seem to suffer from high variance, but little is rigorously known. In this work we study unbiased methods for alpha-divergence minimization through the Signal-to-Noise Ratio (SNR) of the gradient estimator. We study several representative scenarios where strong analytical results are possible, such as fully-factorized or Gaussian distributions. We find that when alpha is not zero, the SNR worsens exponentially in the dimensionality of the problem. This casts doubt on the practicality of these methods. We empirically confirm these theoretical results.
****
How and Why to Use Experimental Data to Evaluate Methods for Observational Causal Inference
Amanda M Gentzel, Purva Pruthi, David Jensen
Methods that infer causal dependence from observational data are central to many areas of science, including medicine, economics, and the social sciences. A variety of theoretical properties of these methods have been proven, but empirical evaluation remains a challenge, largely due to the lack of observational data sets for which treatment effect is known. We describe and analyze observational sampling from randomized controlled trials (OSRCT), a method for evaluating causal inference methods using data from randomized controlled trials (RCTs). This method can be used to create constructed observational data sets with corresponding unbiased estimates of treatment effect, substantially increasing the number of data sets available for evaluating causal inference methods. We show that, in expectation, OSRCT creates data sets that are equivalent to those produced by randomly sampling from empirical data sets in which all potential outcomes are available. We then perform a large-scale evaluation of seven causal inference methods over 37 data sets, drawn from RCTs, as well as simulators, real-world computational systems, and observational data sets augmented with a synthetic response variable. We find notable performance differences when comparing across data from different sources, demonstrating the importance of using data from a variety of sources when evaluating any causal inference method.
****

Strategic Classification in the Dark
Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, Nir Rosenfeld
Strategic classification studies the interaction between a classification rule a
nd the strategic agents it governs. Agents respond by manipulating their feature
s, under the assumption that the classifier is known. However, in many real-life
 scenarios of high-stake classification (e.g., credit scoring), the classifier i
s not revealed to the agents, which leads agents to attempt to learn the classif
ier and game it too. In this paper we generalize the strategic classification mo
del to such scenarios and analyze the effect of an unknown classifier. We define
 the ”price of opacity” as the difference between the prediction error under the
 opaque and transparent policies, characterize it, and give a sufficient conditi
on for it to be strictly positive, in which case transparency is the recommended
 policy. Our experiments show how Hardt et al.’s robust classifier is affected b
y keeping agents in the dark.
****
EMaQ: Expected-Max Q-Learning Operator for Simple Yet Effective Offline and Onli
ne RL
Seyed Kamyar Seyed Ghasemipour, Dale Schuurmans, Shixiang Shane Gu
Off-policy reinforcement learning (RL) holds the promise of sample-efficient lea
rning of decision-making policies by leveraging past experience. However, in the
 offline RL setting – where a fixed collection of interactions are provided and
no further interactions are allowed – it has been shown that standard off-policy
 RL methods can significantly underperform. In this work, we closely investigate
 an important simplification of BCQ (Fujimoto et al., 2018) – a prior approach f
or offline RL – removing a heuristic design choice. Importantly, in contrast to
their original theoretical considerations, we derive this simplified algorithm t
hrough the introduction of a novel backup operator, Expected-Max Q-Learning (EMa
Q), which is more closely related to the resulting practical algorithm. Specific
ally, in addition to the distribution support, EMaQ explicitly considers the num
ber of samples and the proposal distribution, allowing us to derive new sub-opti
mality bounds. In the offline RL setting – the main focus of this work – EMaQ ma
tches and outperforms prior state-of-the-art in the D4RL benchmarks (Fu et al.,
2020). In the online RL setting, we demonstrate that EMaQ is competitive with So
ft Actor Critic (SAC). The key contributions of our empirical findings are demon
strating the importance of careful generative model design for estimating behavi
or policies, and an intuitive notion of complexity for offline RL problems. With
 its simple interpretation and fewer moving parts, such as no explicit function
approximator representing the policy, EMaQ serves as a strong yet easy to implem
ent baseline for future work.
****
Differentially Private Aggregation in the Shuffle Model: Almost Central Accuracy
 in Almost a Single Message
Badih Ghazi, Ravi Kumar, Pasin Manurangsi, Rasmus Pagh, Amer Sinha
The shuffle model of differential privacy has attracted attention in the literat
ure due to it being a middle ground between the well-studied central and local m
odels. In this work, we study the problem of summing (aggregating) real numbers
or integers, a basic primitive in numerous machine learning tasks, in the shuffl
e model. We give a protocol achieving error arbitrarily close to that of the (Di
screte) Laplace mechanism in central differential privacy, while each user only
sends 1 + o(1) short messages in expectation.
****
The Power of Adaptivity for Stochastic Submodular Cover
Rohan Ghuge, Anupam Gupta, Viswanath Nagarajan
In the stochastic submodular cover problem, the goal is to select a subset of st
ochastic items of minimum expected cost to cover a submodular function. Solution
s in this setting correspond to a sequential decision process that selects items
 one by one “adaptively” (depending on prior observations). While such adaptive
solutions achieve the best objective, the inherently sequential nature makes the
m undesirable in many applications. We ask: \emph{how well can solutions with on
ly a few adaptive rounds approximate fully-adaptive solutions?} We consider both

cases where the stochastic items are independent, and where they are correlated. For both situations, we obtain nearly tight answers, establishing smooth tradeoffs between the number of adaptive rounds and the solution quality, relative to fully adaptive solutions. Experiments on synthetic and real datasets validate the practical performance of our algorithms, showing qualitative improvements in the solutions as we allow more rounds of adaptivity; in practice, solutions using just a few rounds of adaptivity are nearly as good as fully adaptive solutions.
****
Differentially Private Quantiles
Jennifer Gillenwater, Matthew Joseph, Alex Kulesza
Quantiles are often used for summarizing and understanding data. If that data is sensitive, it may be necessary to compute quantiles in a way that is differentially private, providing theoretical guarantees that the result does not reveal private information. However, when multiple quantiles are needed, existing differentially private algorithms fare poorly: they either compute quantiles individually, splitting the privacy budget, or summarize the entire distribution, wasting effort. In either case the result is reduced accuracy. In this work we propose an instance of the exponential mechanism that simultaneously estimates exactly m quantiles from n data points while guaranteeing differential privacy. The utility function is carefully structured to allow for an efficient implementation that returns estimates of all m quantiles in time $O(mn\log(n) + m^2n)$. Experiments show that our method significantly outperforms the current state of the art on both real and synthetic data while remaining efficient enough to be practical.
****
Query Complexity of Adversarial Attacks
Grzegorz Gluch, Rüdiger Urbanke
There are two main attack models considered in the adversarial robustness literature: black-box and white-box. We consider these threat models as two ends of a fine-grained spectrum, indexed by the number of queries the adversary can ask. Using this point of view we investigate how many queries the adversary needs to make to design an attack that is comparable to the best possible attack in the white-box model. We give a lower bound on that number of queries in terms of entropy of decision boundaries of the classifier. Using this result we analyze two classical learning algorithms on two synthetic tasks for which we prove meaningful security guarantees. The obtained bounds suggest that some learning algorithms are inherently more robust against query-bounded adversaries than others.
****
Spectral Normalisation for Deep Reinforcement Learning: An Optimisation Perspective
Florin Gogianu, Tudor Berariu, Mihaela C Rosca, Claudia Clopath, Lucian Busoniu, Razvan Pascanu
Most of the recent deep reinforcement learning advances take an RL-centric perspective and focus on refinements of the training objective. We diverge from this view and show we can recover the performance of these developments not by changing the objective, but by regularising the value-function estimator. Constraining the Lipschitz constant of a single layer using spectral normalisation is sufficient to elevate the performance of a Categorical-DQN agent to that of a more elaborated agent on the challenging Atari domain. We conduct ablation studies to disentangle the various effects normalisation has on the learning dynamics and show that is sufficient to modulate the parameter updates to recover most of the performance of spectral normalisation. These findings hint towards the need to also focus on the neural component and its learning dynamics to tackle the peculiarities of Deep Reinforcement Learning.
****
12-Lead ECG Reconstruction via Koopman Operators
Tomer Golany, Kira Radinsky, Daniel Freedman, Saar Minha
32% of all global deaths in the world are caused by cardiovascular diseases. Early detection, especially for patients with ischemia or cardiac arrhythmia, is cr

ucial. To reduce the time between symptoms onset and treatment, wearable ECG sensors were developed to allow for the recording of the full 12-lead ECG signal at home. However, if even a single lead is not correctly positioned on the body that lead becomes corrupted, making automatic diagnosis on the basis of the full signal impossible. In this work, we present a methodology to reconstruct missing or noisy leads using the theory of Koopman Operators. Given a dataset consisting of full 12-lead ECGs, we learn a dynamical system describing the evolution of the 12 individual signals together in time. The Koopman theory indicates that there exists a high-dimensional embedding space in which the operator which propagates from one time instant to the next is linear. We therefore learn both the mapping to this embedding space, as well as the corresponding linear operator. Armed with this representation, we are able to impute missing leads by solving a least squares system in the embedding space, which can be achieved efficiently due to the sparse structure of the system. We perform an empirical evaluation using 12-lead ECG signals from thousands of patients, and show that we are able to reconstruct the signals in such way that enables accurate clinical diagnosis.
****
Function Contrastive Learning of Transferable Meta-Representations
Muhammad Waleed Gondal, Shruti Joshi, Nasim Rahaman, Stefan Bauer, Manuel Wuthrich, Bernhard Schölkopf
Meta-learning algorithms adapt quickly to new tasks that are drawn from the same task distribution as the training tasks. The mechanism leading to fast adaptation is the conditioning of a downstream predictive model on the inferred representation of the task’s underlying data generative process, or \emph{function}. This \emph{meta-representation}, which is computed from a few observed examples of the underlying function, is learned jointly with the predictive model. In this work, we study the implications of this joint training on the transferability of the meta-representations. Our goal is to learn meta-representations that are robust to noise in the data and facilitate solving a wide range of downstream tasks that share the same underlying functions. To this end, we propose a decoupled encoder-decoder approach to supervised meta-learning, where the encoder is trained with a contrastive objective to find a good representation of the underlying function. In particular, our training scheme is driven by the self-supervision signal indicating whether two sets of examples stem from the same function. Our experiments on a number of synthetic and real-world datasets show that the representations we obtain outperform strong baselines in terms of downstream performance and noise robustness, even when these baselines are trained in an end-to-end manner.
****
Active Slices for Sliced Stein Discrepancy
Wenbo Gong, Kaibo Zhang, Yingzhen Li, Jose Miguel Hernandez-Lobato
Sliced Stein discrepancy (SSD) and its kernelized variants have demonstrated promising successes in goodness-of-fit tests and model learning in high dimensions. Despite the theoretical elegance, their empirical performance depends crucially on the search of the optimal slicing directions to discriminate between two distributions. Unfortunately, previous gradient-based optimisation approach returns sub-optimal results for the slicing directions: it is computationally expensive, sensitive to initialization, and it lacks theoretical guarantee for convergence. We address these issues in two steps. First, we show in theory that the requirement of using optimal slicing directions in the kernelized version of SSD can be relaxed, validating the resulting discrepancy with finite random slicing directions. Second, given that good slicing directions are crucial for practical performance, we propose a fast algorithm for finding good slicing directions based on ideas of active sub-space construction and spectral decomposition. Experiments in goodness-of-fit tests and model learning show that our approach achieves both the best performance and the fastest convergence. Especially, we demonstrate 14-80x speed-up in goodness-of-fit tests when compared with the gradient-based approach.
****
On the Problem of Underranking in Group-Fair Ranking

Sruthi Gorantla, Amit Deshpande, Anand Louis
Bias in ranking systems, especially among the top ranks, can worsen social and economic inequalities, polarize opinions, and reinforce stereotypes. On the other hand, a bias correction for minority groups can cause more harm if perceived as favoring group-fair outcomes over meritocracy. Most group-fair ranking algorithms post-process a given ranking and output a group-fair ranking. In this paper, we formulate the problem of underranking in group-fair rankings based on how close the group-fair rank of each item is to its original rank, and prove a lower bound on the trade-off achievable for simultaneous underranking and group fairness in ranking. We give a fair ranking algorithm that takes any given ranking and outputs another ranking with simultaneous underranking and group fairness guarantees comparable to the lower bound we prove. Our experimental results confirm the theoretical trade-off between underranking and group fairness, and also show that our algorithm achieves the best of both when compared to the state-of-the-art baselines.
****
MARINA: Faster Non-Convex Distributed Learning with Compression
Eduard Gorbunov, Konstantin P. Burlachenko, Zhize Li, Peter Richtarik
We develop and analyze MARINA: a new communication efficient method for non-convex distributed learning over heterogeneous datasets. MARINA employs a novel communication compression strategy based on the compression of gradient differences that is reminiscent of but different from the strategy employed in the DIANA method of Mishchenko et al. (2019). Unlike virtually all competing distributed first-order methods, including DIANA, ours is based on a carefully designed biased gradient estimator, which is the key to its superior theoretical and practical performance. The communication complexity bounds we prove for MARINA are evidently better than those of all previous first-order methods. Further, we develop and analyze two variants of MARINA: VR-MARINA and PP-MARINA. The first method is designed for the case when the local loss functions owned by clients are either of a finite sum or of an expectation form, and the second method allows for a partial participation of clients {–} a feature important in federated learning. All our methods are superior to previous state-of-the-art methods in terms of oracle/communication complexity. Finally, we provide a convergence analysis of all methods for problems satisfying the Polyak-{■}ojasiewicz condition.
****
Systematic Analysis of Cluster Similarity Indices: How to Validate Validation Measures
Martijn M Gösgens, Alexey Tikhonov, Liudmila Prokhorenkova
Many cluster similarity indices are used to evaluate clustering algorithms, and choosing the best one for a particular task remains an open problem. We demonstrate that this problem is crucial: there are many disagreements among the indices, these disagreements do affect which algorithms are preferred in applications, and this can lead to degraded performance in real-world systems. We propose a theoretical framework to tackle this problem: we develop a list of desirable properties and conduct an extensive theoretical analysis to verify which indices satisfy them. This allows for making an informed choice: given a particular application, one can first select properties that are desirable for the task and then identify indices satisfying these. Our work unifies and considerably extends existing attempts at analyzing cluster similarity indices: we introduce new properties, formalize existing ones, and mathematically prove or disprove each property for an extensive list of validation indices. This broader and more rigorous approach leads to recommendations that considerably differ from how validation indices are currently being chosen by practitioners. Some of the most popular indices are even shown to be dominated by previously overlooked ones.
****
Revisiting Point Cloud Shape Classification with a Simple and Effective Baseline
Ankit Goyal, Hei Law, Bowei Liu, Alejandro Newell, Jia Deng
Processing point cloud data is an important component of many real-world systems. As such, a wide variety of point-based approaches have been proposed, reporting steady benchmark improvements over time. We study the key ingredients of this

progress and uncover two critical results. First, we find that auxiliary factors
 like different evaluation schemes, data augmentation strategies, and loss funct
ions, which are independent of the model architecture, make a large difference i
n performance. The differences are large enough that they obscure the effect of
architecture. When these factors are controlled for, PointNet++, a relatively ol
der network, performs competitively with recent methods. Second, a very simple p
rojection-based method, which we refer to as SimpleView, performs surprisingly w
ell. It achieves on par or better results than sophisticated state-of-the-art me
thods on ModelNet40 while being half the size of PointNet++. It also outperforms
 state-of-the-art methods on ScanObjectNN, a real-world point cloud benchmark, a
nd demonstrates better cross-dataset generalization. Code is available at https:
//github.com/princeton-vl/SimpleView.
****
Dissecting Supervised Contrastive Learning
Florian Graf, Christoph Hofer, Marc Niethammer, Roland Kwitt
Minimizing cross-entropy over the softmax scores of a linear map composed with a
 high-capacity encoder is arguably the most popular choice for training neural n
etworks on supervised learning tasks. However, recent works show that one can di
rectly optimize the encoder instead, to obtain equally (or even more) discrimina
tive representations via a supervised variant of a contrastive objective. In thi
s work, we address the question whether there are fundamental differences in the
 sought-for representation geometry in the output space of the encoder at minima
l loss. Specifically, we prove, under mild assumptions, that both losses attain
their minimum once the representations of each class collapse to the vertices of
 a regular simplex, inscribed in a hypersphere. We provide empirical evidence th
at this configuration is attained in practice and that reaching a close-to-optim
al state typically indicates good generalization performance. Yet, the two losse
s show remarkably different optimization behavior. The number of iterations requ
ired to perfectly fit to data scales superlinearly with the amount of randomly f
lipped labels for the supervised contrastive loss. This is in contrast to the ap
proximately linear scaling previously reported for networks trained with cross-e
ntropy.
****
Oops I Took A Gradient: Scalable Sampling for Discrete Distributions
Will Grathwohl, Kevin Swersky, Milad Hashemi, David Duvenaud, Chris Maddison
We propose a general and scalable approximate sampling strategy for probabilisti
c models with discrete variables. Our approach uses gradients of the likelihood
function with respect to its discrete inputs to propose updates in a Metropolis-
Hastings sampler. We show empirically that this approach outperforms generic sam
plers in a number of difficult settings including Ising models, Potts models, re
stricted Boltzmann machines, and factorial hidden Markov models. We also demonst
rate our improved sampler for training deep energy-based models on high dimensio
nal discrete image data. This approach outperforms variational auto-encoders and
 existing energy-based models. Finally, we give bounds showing that our approach
 is near-optimal in the class of samplers which propose local updates.
****
Detecting Rewards Deterioration in Episodic Reinforcement Learning
Ido Greenberg, Shie Mannor
In many RL applications, once training ends, it is vital to detect any deteriora
tion in the agent performance as soon as possible. Furthermore, it often has to
be done without modifying the policy and under minimal assumptions regarding the
 environment. In this paper, we address this problem by focusing directly on the
 rewards and testing for degradation. We consider an episodic framework, where t
he rewards within each episode are not independent, nor identically-distributed,
 nor Markov. We present this problem as a multivariate mean-shift detection prob
lem with possibly partial observations. We define the mean-shift in a way corres
ponding to deterioration of a temporal signal (such as the rewards), and derive
a test for this problem with optimal statistical power. Empirically, on deterior
ated rewards in control problems (generated using various environment modificati
ons), the test is demonstrated to be more powerful than standard tests - often b

y orders of magnitude. We also suggest a novel Bootstrap mechanism for False Ala rm Rate control (BFAR), applicable to episodic (non-i.i.d) signal and allowing o ur test to run sequentially in an online manner. Our method does not rely on a l earned model of the environment, is entirely external to the agent, and in fact can be applied to detect changes or drifts in any episodic signal.
****
Crystallization Learning with the Delaunay Triangulation
Jiaqi Gu, Guosheng Yin
Based on the Delaunay triangulation, we propose the crystallization learning to estimate the conditional expectation function in the framework of nonparametric regression. By conducting the crystallization search for the Delaunay simplices closest to the target point in a hierarchical way, the crystallization learning estimates the conditional expectation of the response by fitting a local linear model to the data points of the constructed Delaunay simplices. Instead of condu cting the Delaunay triangulation for the entire feature space which would encoun ter enormous computational difficulty, our approach focuses only on the neighbor hood of the target point and thus greatly expedites the estimation for high-dime nsional cases. Because the volumes of Delaunay simplices are adaptive to the den sity of feature data points, our method selects neighbor data points uniformly i n all directions and thus is more robust to the local geometric structure of the data than existing nonparametric regression methods. We develop the asymptotic properties of the crystallization learning and conduct numerical experiments on both synthetic and real data to demonstrate the advantages of our method in esti mation of the conditional expectation function and prediction of the response.
****
AutoAttend: Automated Attention Representation Search
Chaoyu Guan, Xin Wang, Wenwu Zhu
Self-attention mechanisms have been widely adopted in many machine learning area s, including Natural Language Processing (NLP) and Graph Representation Learning (GRL), etc. However, existing works heavily rely on hand-crafted design to obta in customized attention mechanisms. In this paper, we automate Key, Query and Va lue representation design, which is one of the most important steps to obtain ef fective self-attentions. We propose an automated self-attention representation m odel, AutoAttend, which can automatically search powerful attention representati ons for downstream tasks leveraging Neural Architecture Search (NAS). In particu lar, we design a tailored search space for attention representation automation, which is flexible to produce effective attention representation designs. Based o n the design prior obtained from attention representations in previous works, we further regularize our search space to reduce the space complexity without the loss of expressivity. Moreover, we propose a novel context-aware parameter shari ng mechanism considering special characteristics of each sub-architecture to pro vide more accurate architecture estimations when conducting parameter sharing in our tailored search space. Experiments show the superiority of our proposed Aut oAttend model over previous state-of-the-arts on eight text classification tasks in NLP and four node classification tasks in GRL.
****
Operationalizing Complex Causes: A Pragmatic View of Mediation
Limor Gultchin, David Watson, Matt Kusner, Ricardo Silva
We examine the problem of causal response estimation for complex objects (e.g., text, images, genomics). In this setting, classical \emph{atomic} interventions are often not available (e.g., changes to characters, pixels, DNA base-pairs). I nstead, we only have access to indirect or \emph{crude} interventions (e.g., enr olling in a writing program, modifying a scene, applying a gene therapy). In thi s work, we formalize this problem and provide an initial solution. Given a colle ction of candidate mediators, we propose (a) a two-step method for predicting th e causal responses of crude interventions; and (b) a testing procedure to identi fy mediators of crude interventions. We demonstrate, on a range of simulated and real-world-inspired examples, that our approach allows us to efficiently estima te the effect of crude interventions with limited data from new treatment regime s.

****
On a Combination of Alternating Minimization and Nesterov’s Momentum
Sergey Guminov, Pavel Dvurechensky, Nazarii Tupitsa, Alexander Gasnikov
Alternating minimization (AM) procedures are practically efficient in many applications for solving convex and non-convex optimization problems. On the other hand, Nesterov’s accelerated gradient is theoretically optimal first-order method for convex optimization. In this paper we combine AM and Nesterov’s acceleration to propose an accelerated alternating minimization algorithm. We prove $1/k^2$ convergence rate in terms of the objective for convex problems and $1/k$ in terms of the squared gradient norm for non-convex problems, where k is the iteration counter. Our method does not require any knowledge of neither convexity of the problem nor function parameters such as Lipschitz constant of the gradient, i.e. it is adaptive to convexity and smoothness and is uniformly optimal for smooth convex and non-convex problems. Further, we develop its primal-dual modification for strongly convex problems with linear constraints and prove the same $1/k^2$ for the primal objective residual and constraints feasibility.
****
Decentralized Single-Timescale Actor-Critic on Zero-Sum Two-Player Stochastic Games
Hongyi Guo, Zuyue Fu, Zhuoran Yang, Zhaoran Wang
We study the global convergence and global optimality of the actor-critic algorithm applied for the zero-sum two-player stochastic games in a decentralized manner. We focus on the single-timescale setting where the critic is updated by applying the Bellman operator only once and the actor is updated by policy gradient with the information from the critic. Our algorithm is in a decentralized manner, as we assume that each player has no access to the actions of the other one, which, in a way, protects the privacy of both players. Moreover, we consider linear function approximations for both actor and critic, and we prove that the sequence of joint policy generated by our decentralized linear algorithm converges to the minimax equilibrium at a sublinear rate \(\cO(\sqrt{K})\), where \(K\){is} the number of iterations. To the best of our knowledge, we establish the global optimality and convergence of decentralized actor-critic algorithm on zero-sum two-player stochastic games with linear function approximations for the first time.
****
Adversarial Policy Learning in Two-player Competitive Games
Wenbo Guo, Xian Wu, Sui Huang, Xinyu Xing
In a two-player deep reinforcement learning task, recent work shows an attacker could learn an adversarial policy that triggers a target agent to perform poorly and even react in an undesired way. However, its efficacy heavily relies upon the zero-sum assumption made in the two-player game. In this work, we propose a new adversarial learning algorithm. It addresses the problem by resetting the optimization goal in the learning process and designing a new surrogate optimization function. Our experiments show that our method significantly improves adversarial agents’ exploitability compared with the state-of-art attack. Besides, we also discover that our method could augment an agent with the ability to abuse the target game’s unfairness. Finally, we show that agents adversarially re-trained against our adversarial agents could obtain stronger adversary-resistance.
****
Soft then Hard: Rethinking the Quantization in Neural Image Compression
Zongyu Guo, Zhizheng Zhang, Runsen Feng, Zhibo Chen
Quantization is one of the core components in lossy image compression. For neural image compression, end-to-end optimization requires differentiable approximations of quantization, which can generally be grouped into three categories: additive uniform noise, straight-through estimator and soft-to-hard annealing. Training with additive uniform noise approximates the quantization error variationally but suffers from the train-test mismatch. The other two methods do not encounter this mismatch but, as shown in this paper, hurt the rate-distortion performance since the latent representation ability is weakened. We thus propose a novel soft-then-hard quantization strategy for neural image compression that first lear

ns an expressive latent space softly, then closes the train-test mismatch with hard quantization. In addition, beyond the fixed integer-quantization, we apply scaled additive uniform noise to adaptively control the quantization granularity by deriving a new variational upper bound on actual rate. Experiments demonstrate that our proposed methods are easy to adopt, stable to train, and highly effective especially on complex compression models.
****
UneVEn: Universal Value Exploration for Multi-Agent Reinforcement Learning
Tarun Gupta, Anuj Mahajan, Bei Peng, Wendelin Boehmer, Shimon Whiteson
VDN and QMIX are two popular value-based algorithms for cooperative MARL that learn a centralized action value function as a monotonic mixing of per-agent utilities. While this enables easy decentralization of the learned policy, the restricted joint action value function can prevent them from solving tasks that require significant coordination between agents at a given timestep. We show that this problem can be overcome by improving the joint exploration of all agents during training. Specifically, we propose a novel MARL approach called Universal Value Exploration (UneVEn) that learns a set of related tasks simultaneously with a linear decomposition of universal successor features. With the policies of already solved related tasks, the joint exploration process of all agents can be improved to help them achieve better coordination. Empirical results on a set of exploration games, challenging cooperative predator-prey tasks requiring significant coordination among agents, and StarCraft II micromanagement benchmarks show that UneVEn can solve tasks where other state-of-the-art MARL methods fail.
****
Distribution-Free Calibration Guarantees for Histogram Binning without Sample Splitting
Chirag Gupta, Aaditya Ramdas
We prove calibration guarantees for the popular histogram binning (also called uniform-mass binning) method of Zadrozny and Elkan (2001). Histogram binning has displayed strong practical performance, but theoretical guarantees have only been shown for sample split versions that avoid ’double dipping’ the data. We demonstrate that the statistical cost of sample splitting is practically significant on a credit default dataset. We then prove calibration guarantees for the original method that double dips the data, using a certain Markov property of order statistics. Based on our results, we make practical recommendations for choosing the number of bins in histogram binning. In our illustrative simulations, we propose a new tool for assessing calibration—validity plots—which provide more information than an ECE estimate.
****
Correcting Exposure Bias for Link Recommendation
Shantanu Gupta, Hao Wang, Zachary Lipton, Yuyang Wang
Link prediction methods are frequently applied in recommender systems, e.g., to suggest citations for academic papers or friends in social networks. However, exposure bias can arise when users are systematically underexposed to certain relevant items. For example, in citation networks, authors might be more likely to encounter papers from their own field and thus cite them preferentially. This bias can propagate through naively trained link predictors, leading to both biased evaluation and high generalization error (as assessed by true relevance). Moreover, this bias can be exacerbated by feedback loops. We propose estimators that leverage known exposure probabilities to mitigate this bias and consequent feedback loops. Next, we provide a loss function for learning the exposure probabilities from data. Finally, experiments on semi-synthetic data based on real-world citation networks, show that our methods reliably identify (truly) relevant citations. Additionally, our methods lead to greater diversity in the recommended papers’ fields of study. The code is available at github.com/shantanu95/exposure-bias-link-rec.
****
The Heavy-Tail Phenomenon in SGD
Mert Gurbuzbalaban, Umut Simsekli, Lingjiong Zhu
In recent years, various notions of capacity and complexity have been proposed f

or characterizing the generalization properties of stochastic gradient descent (SGD) in deep learning. Some of the popular notions that correlate well with the performance on unseen data are (i) the ‘flatness’ of the local minimum found by SGD, which is related to the eigenvalues of the Hessian, (ii) the ratio of the stepsize η to the batch-size b, which essentially controls the magnitude of the stochastic gradient noise, and (iii) the ‘tail-index’, which measures the heaviness of the tails of the network weights at convergence. In this paper, we argue that these three seemingly unrelated perspectives for generalization are deeply linked to each other. We claim that depending on the structure of the Hessian of the loss at the minimum, and the choices of the algorithm parameters η and b, the SGD iterates will converge to a \emph{heavy-tailed} stationary distribution. We rigorously prove this claim in the setting of quadratic optimization: we show that even in a simple linear regression problem with independent and identically distributed data whose distribution has finite moments of all order, the iterates can be heavy-tailed with infinite variance. We further characterize the behavior of the tails with respect to algorithm parameters, the dimension, and the curvature. We then translate our results into insights about the behavior of SGD in deep learning. We support our theory with experiments conducted on synthetic data, fully connected, and convolutional neural networks.
****
Knowledge Enhanced Machine Learning Pipeline against Diverse Adversarial Attacks
Nezihe Merve Gürel, Xiangyu Qi, Luka Rimanic, Ce Zhang, Bo Li
Despite the great successes achieved by deep neural networks (DNNs), recent studies show that they are vulnerable against adversarial examples, which aim to mislead DNNs by adding small adversarial perturbations. Several defenses have been proposed against such attacks, while many of them have been adaptively attacked. In this work, we aim to enhance the ML robustness from a different perspective by leveraging domain knowledge: We propose a Knowledge Enhanced Machine Learning Pipeline (KEMLP) to integrate domain knowledge (i.e., logic relationships among different predictions) into a probabilistic graphical model via first-order logic rules. In particular, we develop KEMLP by integrating a diverse set of weak auxiliary models based on their logical relationships to the main DNN model that performs the target task. Theoretically, we provide convergence results and prove that, under mild conditions, the prediction of KEMLP is more robust than that of the main DNN model. Empirically, we take road sign recognition as an example and leverage the relationships between road signs and their shapes and contents as domain knowledge. We show that compared with adversarial training and other baselines, KEMLP achieves higher robustness against physical attacks, $\mathcal{L}_p$ bounded attacks, unforeseen attacks, and natural corruptions under both whitebox and blackbox settings, while still maintaining high clean accuracy.
****
Adapting to Delays and Data in Adversarial Multi-Armed Bandits
Andras Gyorgy, Pooria Joulani
We consider the adversarial multi-armed bandit problem under delayed feedback. We analyze variants of the Exp3 algorithm that tune their step size using only information (about the losses and delays) available at the time of the decisions, and obtain regret guarantees that adapt to the observed (rather than the worst-case) sequences of delays and/or losses. First, through a remarkably simple proof technique, we show that with proper tuning of the step size, the algorithm achieves an optimal (up to logarithmic factors) regret of order $\sqrt{\log(K)(TK + D)}$ both in expectation and in high probability, where K is the number of arms, T is the time horizon, and D is the cumulative delay. The high-probability version of the bound, which is the first high-probability delay-adaptive bound in the literature, crucially depends on the use of implicit exploration in estimating the losses. Then, following Zimmert and Seldin (2019), we extend these results so that the algorithm can “skip” rounds with large delays, resulting in regret bounds of order $\sqrt{TK\log(K)} + |R| + \sqrt{D_{\bar{R}}\log(K)}$, where R is an arbitrary set of rounds (which are skipped) and $D_{\bar{R}}$ is the cumulative delay of the feedback for other rounds. Finally, we present another, data-adaptive (AdaGrad-style) version of the algorithm for which the regret adap

ts to the observed (delayed) losses instead of only adapting to the cumulative delay (this algorithm requires an a priori upper bound on the maximum delay, or the advance knowledge of the delay for each decision when it is made). The resulting bound can be orders of magnitude smaller on benign problems, and it can be shown that the delay only affects the regret through the loss of the best arm.
****
Rate-Distortion Analysis of Minimum Excess Risk in Bayesian Learning
Hassan Hafez-Kolahi, Behrad Moniri, Shohreh Kasaei, Mahdieh Soleymani Baghshah
In parametric Bayesian learning, a prior is assumed on the parameter W which determines the distribution of samples. In this setting, Minimum Excess Risk (MER) is defined as the difference between the minimum expected loss achievable when learning from data and the minimum expected loss that could be achieved if W was observed. In this paper, we build upon and extend the recent results of (Xu & Raginsky, 2020) to analyze the MER in Bayesian learning and derive information-theoretic bounds on it. We formulate the problem as a (constrained) rate-distortion optimization and show how the solution can be bounded above and below by two other rate-distortion functions that are easier to study. The lower bound represents the minimum possible excess risk achievable by \emph{any} process using R bits of information from the parameter W. For the upper bound, the optimization is further constrained to use R bits from the training set, a setting which relates MER to information-theoretic bounds on the generalization gap in frequentist learning. We derive information-theoretic bounds on the difference between these upper and lower bounds and show that they can provide order-wise tight rates for MER under certain conditions. This analysis gives more insight into the information-theoretic nature of Bayesian learning as well as providing novel bounds.
****
Regret Minimization in Stochastic Non-Convex Learning via a Proximal-Gradient Approach
Nadav Hallak, Panayotis Mertikopoulos, Volkan Cevher
This paper develops a methodology for regret minimization with stochastic first-order oracle feedback in online, constrained, non-smooth, non-convex problems. In this setting, the minimization of external regret is beyond reach for first-order methods, and there are no gradient-based algorithmic frameworks capable of providing a solution. On that account, we propose a conceptual approach that leverages non-convex optimality measures, leading to a suitable generalization of the learner’s local regret. We focus on a local regret measure defined via a proximal-gradient mapping, that also encompasses the original notion proposed by Hazan et al. (2017). To achieve no local regret in this setting, we develop a proximal-gradient method based on stochastic first-order feedback, and a simpler method for when access to a perfect first-order oracle is possible. Both methods are order-optimal (in the min-max sense), and we also establish a bound on the number of proximal-gradient queries these methods require. As an important application of our results, we also obtain a link between online and offline non-convex stochastic optimization manifested as a new proximal-gradient scheme with complexity guarantees matching those obtained via variance reduction techniques.
****
Diversity Actor-Critic: Sample-Aware Entropy Regularization for Sample-Efficient Exploration
Seungyul Han, Youngchul Sung
In this paper, sample-aware policy entropy regularization is proposed to enhance the conventional policy entropy regularization for better exploration. Exploiting the sample distribution obtainable from the replay buffer, the proposed sample-aware entropy regularization maximizes the entropy of the weighted sum of the policy action distribution and the sample action distribution from the replay buffer for sample-efficient exploration. A practical algorithm named diversity actor-critic (DAC) is developed by applying policy iteration to the objective function with the proposed sample-aware entropy regularization. Numerical results show that DAC significantly outperforms existing recent algorithms for reinforcement learning.

****
Adversarial Combinatorial Bandits with General Non-linear Reward Functions
Yanjun Han, Yining Wang, Xi Chen
In this paper we study the adversarial combinatorial bandit with a known non-linear reward function, extending existing work on adversarial linear combinatorial bandit. {The adversarial combinatorial bandit with general non-linear reward is an important open problem in bandit literature, and it is still unclear whether there is a significant gap from the case of linear reward, stochastic bandit, or semi-bandit feedback.} We show that, with N arms and subsets of K arms being chosen at each of T time periods, the minimax optimal regret is $\widetilde\Theta_{d}(\sqrt{N^d T})$ if the reward function is a d-degree polynomial with $d< K$, and $\Theta_K(\sqrt{N^K T})$ if the reward function is not a low-degree polynomial. {Both bounds are significantly different from the bound $O(\sqrt{\mathrm{poly}(N,K)T})$ for the linear case, which suggests that there is a fundamental gap between the linear and non-linear reward structures.} Our result also finds applications to adversarial assortment optimization problem in online recommendation. We show that in the worst-case of adversarial assortment problem, the optimal algorithm must treat each individual $\binom{N}{K}$ assortment as independent.
****
A Collective Learning Framework to Boost GNN Expressiveness for Node Classification
Mengyue Hang, Jennifer Neville, Bruno Ribeiro
Collective Inference (CI) is a procedure designed to boost weak relational classifiers, specially for node classification tasks. Graph Neural Networks (GNNs) are strong classifiers that have been used with great success. Unfortunately, most existing practical GNNs are not most-expressive (universal). Thus, it is an open question whether one can improve strong relational node classifiers, such as GNNs, with CI. In this work, we investigate this question and propose {\em collective learning} for GNNs —a general collective classification approach for node representation learning that increases their representation power. We show that previous attempts to incorporate CI into GNNs fail to boost their expressiveness because they do not adapt CI’s Monte Carlo sampling to representation learning. We evaluate our proposed framework with a variety of state-of-the-art GNNs. Our experiments show a consistent, significant boost in node classification accuracy —regardless of the choice of underlying GNN— for inductive node classification in partially-labeled graphs, across five real-world network datasets.
****
Grounding Language to Entities and Dynamics for Generalization in Reinforcement Learning
Austin W. Hanjie, Victor Y Zhong, Karthik Narasimhan
We investigate the use of natural language to drive the generalization of control policies and introduce the new multi-task environment Messenger with free-form text manuals describing the environment dynamics. Unlike previous work, Messenger does not assume prior knowledge connecting text and state observations {—} the control policy must simultaneously ground the game manual to entity symbols and dynamics in the environment. We develop a new model, EMMA (Entity Mapper with Multi-modal Attention) which uses an entity-conditioned attention module that allows for selective focus over relevant descriptions in the manual for each entity in the environment. EMMA is end-to-end differentiable and learns a latent grounding of entities and dynamics from text to observations using only environment rewards. EMMA achieves successful zero-shot generalization to unseen games with new dynamics, obtaining a 40% higher win rate compared to multiple baselines. However, win rate on the hardest stage of Messenger remains low (10%), demonstrating the need for additional work in this direction.
****
Sparse Feature Selection Makes Batch Reinforcement Learning More Sample Efficient
Botao Hao, Yaqi Duan, Tor Lattimore, Csaba Szepesvari, Mengdi Wang
This paper provides a statistical analysis of high-dimensional batch reinforceme

nt learning (RL) using sparse linear function approximation. When there is a large number of candidate features, our result sheds light on the fact that sparsity-aware methods can make batch RL more sample efficient. We first consider the off-policy policy evaluation problem. To evaluate a new target policy, we analyze a Lasso fitted Q-evaluation method and establish a finite-sample error bound that has no polynomial dependence on the ambient dimension. To reduce the Lasso bias, we further propose a post model-selection estimator that applies fitted Q-evaluation to the features selected via group Lasso. Under an additional signal strength assumption, we derive a sharper instance-dependent error bound that depends on a divergence function measuring the distribution mismatch between the data distribution and occupancy measure of the target policy. Further, we study the Lasso fitted Q-iteration for batch policy optimization and establish a finite-sample error bound depending on the ratio between the number of relevant features and restricted minimal eigenvalue of the data's covariance. In the end, we complement the results with minimax lower bounds for batch-data policy evaluation/optimization that nearly match our upper bounds. The results suggest that having well-conditioned data is crucial for sparse batch policy learning.
****
Bootstrapping Fitted Q-Evaluation for Off-Policy Inference
Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvari, Mengdi Wang
Bootstrapping provides a flexible and effective approach for assessing the quality of batch reinforcement learning, yet its theoretical properties are poorly understood. In this paper, we study the use of bootstrapping in off-policy evaluation (OPE), and in particular, we focus on the fitted Q-evaluation (FQE) that is known to be minimax-optimal in the tabular and linear-model cases. We propose a bootstrapping FQE method for inferring the distribution of the policy evaluation error and show that this method is asymptotically efficient and distributionally consistent for off-policy statistical inference. To overcome the computation limit of bootstrapping, we further adapt a subsampling procedure that improves the runtime by an order of magnitude. We numerically evaluate the bootrapping method in classical RL environments for confidence interval estimation, estimating the variance of off-policy evaluator, and estimating the correlation between multiple off-policy evaluators.
****
Compressed Maximum Likelihood
Yi Hao, Alon Orlitsky
Maximum likelihood (ML) is one of the most fundamental and general statistical estimation techniques. Inspired by recent advances in estimating distribution functionals, we propose $\textit{compressed maximum likelihood}$ (CML) that applies ML to the compressed samples. We then show that CML is sample-efficient for several essential learning tasks over both discrete and continuous domains, including learning densities with structures, estimating probability multisets, and inferring symmetric distribution functionals.
****
Valid Causal Inference with (Some) Invalid Instruments
Jason S Hartford, Victor Veitch, Dhanya Sridhar, Kevin Leyton-Brown
Instrumental variable methods provide a powerful approach to estimating causal effects in the presence of unobserved confounding. But a key challenge when applying them is the reliance on untestable "exclusion" assumptions that rule out any relationship between the instrument variable and the response that is not mediated by the treatment. In this paper, we show how to perform consistent IV estimation despite violations of the exclusion assumption. In particular, we show that when one has multiple candidate instruments, only a majority of these candidates—or, more generally, the modal candidate-response relationship—needs to be valid to estimate the causal effect. Our approach uses an estimate of the modal prediction from an ensemble of instrumental variable estimators. The technique is simple to apply and is "black-box" in the sense that it may be used with any instrumental variable estimator as long as the treatment effect is identified for each valid instrument independently. As such, it is compatible with recent machine-learning based estimators that allow for the estimation of conditional average t

reatment effects (CATE) on complex, high dimensional data. Experimentally, we achieve accurate estimates of conditional average treatment effects using an ensemble of deep network-based estimators, including on a challenging simulated Mendelian Randomization problem.
****
Model Performance Scaling with Multiple Data Sources
Tatsunori Hashimoto
Real-world machine learning systems are often trained using a mix of data sources with varying cost and quality. Understanding how the size and composition of a training dataset affect model performance is critical for advancing our understanding of generalization, as well as designing more effective data collection policies. We show that there is a simple scaling law that predicts the loss incurred by a model even under varying dataset composition. Our work expands recent observations of scaling laws for log-linear generalization error in the i.i.d setting and uses this to cast model performance prediction as a learning problem. Using the theory of optimal experimental design, we derive a simple rational function approximation to generalization error that can be fitted using a few model training runs. Our approach can achieve highly accurate ($r^2\approx .9$) predictions of model performance under substantial extrapolation in two different standard supervised learning tasks and is accurate ($r^2 \approx .83$) on more challenging machine translation and question answering tasks where many baselines achieve worse-than-random performance.
****
Hierarchical VAEs Know What They Don’t Know
Jakob D. Havtorn, Jes Frellsen, Søren Hauberg, Lars Maaløe
Deep generative models have been demonstrated as state-of-the-art density estimators. Yet, recent work has found that they often assign a higher likelihood to data from outside the training distribution. This seemingly paradoxical behavior has caused concerns over the quality of the attained density estimates. In the context of hierarchical variational autoencoders, we provide evidence to explain this behavior by out-of-distribution data having in-distribution low-level features. We argue that this is both expected and desirable behavior. With this insight in hand, we develop a fast, scalable and fully unsupervised likelihood-ratio score for OOD detection that requires data to be in-distribution across all feature-levels. We benchmark the method on a vast set of data and model combinations and achieve state-of-the-art results on out-of-distribution detection.
****
SPECTRE: defending against backdoor attacks using robust statistics
Jonathan Hayase, Weihao Kong, Raghav Somani, Sewoong Oh
Modern machine learning increasingly requires training on a large collection of data from multiple sources, not all of which can be trusted. A particularly frightening scenario is when a small fraction of corrupted data changes the behavior of the trained model when triggered by an attacker-specified watermark. Such a compromised model will be deployed unnoticed as the model is accurate otherwise. There has been promising attempts to use the intermediate representations of such a model to separate corrupted examples from clean ones. However, these methods require a significant fraction of the data to be corrupted, in order to have strong enough signal for detection. We propose a novel defense algorithm using robust covariance estimation to amplify the spectral signature of corrupted data. This defense is able to completely remove backdoors whenever the benchmark backdoor attacks are successful, even in regimes where previous methods have no hope for detecting poisoned examples.
****
Boosting for Online Convex Optimization
Elad Hazan, Karan Singh
We consider the decision-making framework of online convex optimization with a very large number of experts. This setting is ubiquitous in contextual and reinforcement learning problems, where the size of the policy class renders enumeration and search within the policy class infeasible. Instead, we consider generalizing the methodology of online boosting. We define a weak learning algorithm as a

mechanism that guarantees multiplicatively approximate regret against a base class of experts. In this access model, we give an efficient boosting algorithm that guarantees near-optimal regret against the convex hull of the base class. We consider both full and partial (a.k.a. bandit) information feedback models. We also give an analogous efficient boosting algorithm for the i.i.d. statistical setting. Our results simultaneously generalize online boosting and gradient boosting guarantees to contextual learning model, online convex optimization and bandit linear optimization settings.
****
PipeTransformer: Automated Elastic Pipelining for Distributed Training of Large-scale Models
Chaoyang He, Shen Li, Mahdi Soltanolkotabi, Salman Avestimehr
The size of Transformer models is growing at an unprecedented rate. It has taken less than one year to reach trillion-level parameters since the release of GPT-3 (175B). Training such models requires both substantial engineering efforts and enormous computing resources, which are luxuries most research teams cannot afford. In this paper, we propose PipeTransformer, which leverages automated elastic pipelining for efficient distributed training of Transformer models. In PipeTransformer, we design an adaptive on the fly freeze algorithm that can identify and freeze some layers gradually during training, and an elastic pipelining system that can dynamically allocate resources to train the remaining active layers. More specifically, PipeTransformer automatically excludes frozen layers from the pipeline, packs active layers into fewer GPUs, and forks more replicas to increase data-parallel width. We evaluate PipeTransformer using Vision Transformer (ViT) on ImageNet and BERT on SQuAD and GLUE datasets. Our results show that compared to the state-of-the-art baseline, PipeTransformer attains up to 2.83-fold speedup without losing accuracy. We also provide various performance analyses for a more comprehensive understanding of our algorithmic and system-wise design. Finally, we have modularized our training system with flexible APIs and made the source code publicly available at https://DistML.ai.
****
SoundDet: Polyphonic Moving Sound Event Detection and Localization from Raw Waveform
Yuhang He, Niki Trigoni, Andrew Markham
We present a new framework SoundDet, which is an end-to-end trainable and light-weight framework, for polyphonic moving sound event detection and localization. Prior methods typically approach this problem by preprocessing raw waveform into time-frequency representations, which is more amenable to process with well-established image processing pipelines. Prior methods also detect in segment-wise manner, leading to incomplete and partial detections. SoundDet takes a novel approach and directly consumes the raw, multichannel waveform and treats the spatio-temporal sound event as a complete “sound-object" to be detected. Specifically, SoundDet consists of a backbone neural network and two parallel heads for temporal detection and spatial localization, respectively. Given the large sampling rate of raw waveform, the backbone network first learns a set of phase-sensitive and frequency-selective bank of filters to explicitly retain direction-of-arrival information, whilst being highly computationally and parametrically efficient than standard 1D/2D convolution. A dense sound event proposal map is then constructed to handle the challenges of predicting events with large varying temporal duration. Accompanying the dense proposal map are a temporal overlapness map and a motion smoothness map that measure a proposal’s confidence to be an event from temporal detection accuracy and movement consistency perspective. Involving the two maps guarantees SoundDet to be trained in a spatio-temporally unified manner. Experimental results on the public DCASE dataset show the advantage of SoundDet on both segment-based evaluation and our newly proposed event-based evaluation system.
****
Logarithmic Regret for Reinforcement Learning with Linear Function Approximation
Jiafan He, Dongruo Zhou, Quanquan Gu
Reinforcement learning (RL) with linear function approximation has received incr

easing attention recently. However, existing work has focused on obtaining $\sqrt{T}$-type regret bound, where T is the number of interactions with the MDP. In this paper, we show that logarithmic regret is attainable under two recently proposed linear MDP assumptions provided that there exists a positive sub-optimality gap for the optimal action-value function. More specifically, under the linear MDP assumption (Jin et al., 2020), the LSVI-UCB algorithm can achieve $\tilde{O}(d^{3}H^5/\text{gap}_{\text{min}}\cdot \log(T))$regret; and under the linear mixture MDP assumption (Ayoub et al., 2020), the UCRL-VTR algorithm can achieve $\tilde{O}(d^{2}H^5/\text{gap}_{\text{min}}\cdot \log^3(T))$ regret, where d is the dimension of feature mapping, H is the length of episode, gap_{min} is the minimal sub-optimality gap, and $\tilde O$ hides all logarithmic terms except $\log(T)$. To the best of our knowledge, these are the first logarithmic regret bounds for RL with linear function approximation. We also establish gap-dependent lower bounds for the two linear MDP models.
****
Finding Relevant Information via a Discrete Fourier Expansion
Mohsen Heidari, Jithin Sreedharan, Gil I Shamir, Wojciech Szpankowski
A fundamental obstacle in learning information from data is the presence of nonlinear redundancies and dependencies in it. To address this, we propose a Fourier-based approach to extract relevant information in the supervised setting. We first develop a novel Fourier expansion for functions of correlated binary random variables. This expansion is a generalization of the standard Fourier analysis on the Boolean cube beyond product probability spaces. We further extend our Fourier analysis to stochastic mappings. As an important application of this analysis, we investigate learning with feature subset selection. We reformulate this problem in the Fourier domain and introduce a computationally efficient measure for selecting features. Bridging the Bayesian error rate with the Fourier coefficients, we demonstrate that the Fourier expansion provides a powerful tool to characterize nonlinear dependencies in the features-label relation. Via theoretical analysis, we show that our proposed measure finds provably asymptotically optimal feature subsets. Lastly, we present an algorithm based on our measure and verify our findings via numerical experiments on various datasets.
****
Zeroth-Order Non-Convex Learning via Hierarchical Dual Averaging
Amélie Héliou, Matthieu Martin, Panayotis Mertikopoulos, Thibaud Rahier
We propose a hierarchical version of dual averaging for zeroth-order online non-convex optimization {–} i.e., learning processes where, at each stage, the optimizer is facing an unknown non-convex loss function and only receives the incurred loss as feedback. The proposed class of policies relies on the construction of an online model that aggregates loss information as it arrives, and it consists of two principal components: (a) a regularizer adapted to the Fisher information metric (as opposed to the metric norm of the ambient space); and (b) a principled exploration of the problem’s state space based on an adapted hierarchical schedule. This construction enables sharper control of the model’s bias and variance, and allows us to derive tight bounds for both the learner’s static and dynamic regret {–} i.e., the regret incurred against the best dynamic policy in hindsight over the horizon of play.
****
Improving Molecular Graph Neural Network Explainability with Orthonormalization and Induced Sparsity
Ryan Henderson, Djork-Arné Clevert, Floriane Montanari
Rationalizing which parts of a molecule drive the predictions of a molecular graph convolutional neural network (GCNN) can be difficult. To help, we propose two simple regularization techniques to apply during the training of GCNNs: Batch Representation Orthonormalization (BRO) and Gini regularization. BRO, inspired by molecular orbital theory, encourages graph convolution operations to generate orthonormal node embeddings. Gini regularization is applied to the weights of the output layer and constrains the number of dimensions the model can use to make predictions. We show that Gini and BRO regularization can improve the accuracy of state-of-the-art GCNN attribution methods on artificial benchmark datasets. In

a real-world setting, we demonstrate that medicinal chemists significantly prefer explanations extracted from regularized models. While we only study these regularizers in the context of GCNNs, both can be applied to other types of neural networks.
****
Muesli: Combining Improvements in Policy Optimization
Matteo Hessel, Ivo Danihelka, Fabio Viola, Arthur Guez, Simon Schmitt, Laurent Sifre, Theophane Weber, David Silver, Hado Van Hasselt
We propose a novel policy update that combines regularized policy optimization with model learning as an auxiliary loss. The update (henceforth Muesli) matches MuZero’s state-of-the-art performance on Atari. Notably, Muesli does so without using deep search: it acts directly with a policy network and has computation speed comparable to model-free baselines. The Atari results are complemented by extensive ablations, and by additional results on continuous control and 9x9 Go.
****
Learning Representations by Humans, for Humans
Sophie Hilgard, Nir Rosenfeld, Mahzarin R Banaji, Jack Cao, David Parkes
When machine predictors can achieve higher performance than the human decision-makers they support, improving the performance of human decision-makers is often conflated with improving machine accuracy. Here we propose a framework to directly support human decision-making, in which the role of machines is to reframe problems rather than to prescribe actions through prediction. Inspired by the success of representation learning in improving performance of machine predictors, our framework learns human-facing representations optimized for human performance. This “Mind Composed with Machine” framework incorporates a human decision-making model directly into the representation learning paradigm and is trained with a novel human-in-the-loop training procedure. We empirically demonstrate the successful application of the framework to various tasks and representational forms.
****
Optimizing Black-box Metrics with Iterative Example Weighting
Gaurush Hiranandani, Jatin Mathur, Harikrishna Narasimhan, Mahdi Milani Fard, Sanmi Koyejo
We consider learning to optimize a classification metric defined by a black-box function of the confusion matrix. Such black-box learning settings are ubiquitous, for example, when the learner only has query access to the metric of interest, or in noisy-label and domain adaptation applications where the learner must evaluate the metric via performance evaluation using a small validation sample. Our approach is to adaptively learn example weights on the training dataset such that the resulting weighted objective best approximates the metric on the validation sample. We show how to model and estimate the example weights and use them to iteratively post-shift a pre-trained class probability estimator to construct a classifier. We also analyze the resulting procedure’s statistical properties. Experiments on various label noise, domain shift, and fair classification setups confirm that our proposal compares favorably to the state-of-the-art baselines for each application.
****
Trees with Attention for Set Prediction Tasks
Roy Hirsch, Ran Gilad-Bachrach
In many machine learning applications, each record represents a set of items. For example, when making predictions from medical records, the medications prescribed to a patient are a set whose size is not fixed and whose order is arbitrary. However, most machine learning algorithms are not designed to handle set structures and are limited to processing records of fixed size. Set-Tree, presented in this work, extends the support for sets to tree-based models, such as Random-Forest and Gradient-Boosting, by introducing an attention mechanism and set-compatible split criteria. We evaluate the new method empirically on a wide range of problems ranging from making predictions on sub-atomic particle jets to estimating the redshift of galaxies. The new method outperforms existing tree-based methods consistently and significantly. Moreover, it is competitive and often outperf

orms Deep Learning. We also discuss the theoretical properties of Set-Trees and explain how they enable item-level explainability.
****

Multiplicative Noise and Heavy Tails in Stochastic Optimization
Liam Hodgkinson, Michael Mahoney
Although stochastic optimization is central to modern machine learning, the precise mechanisms underlying its success, and in particular, the precise role of the stochasticity, still remain unclear. Modeling stochastic optimization algorithms as discrete random recurrence relations, we show that multiplicative noise, as it commonly arises due to variance in local rates of convergence, results in heavy-tailed stationary behaviour in the parameters. Theoretical results are obtained characterizing this for a large class of (non-linear and even non-convex) models and optimizers (including momentum, Adam, and stochastic Newton), demonstrating that this phenomenon holds generally. We describe dependence on key factors, including step size, batch size, and data variability, all of which exhibit similar qualitative behavior to recent empirical results on state-of-the-art neural network models. Furthermore, we empirically illustrate how multiplicative noise and heavy-tailed structure improve capacity for basin hopping and exploration of non-convex loss surfaces, over commonly-considered stochastic dynamics with only additive noise and light-tailed structure.
****

MC-LSTM: Mass-Conserving LSTM
Pieter-Jan Hoedt, Frederik Kratzert, Daniel Klotz, Christina Halmich, Markus Holzleitner, Grey S Nearing, Sepp Hochreiter, Guenter Klambauer
The success of Convolutional Neural Networks (CNNs) in computer vision is mainly driven by their strong inductive bias, which is strong enough to allow CNNs to solve vision-related tasks with random weights, meaning without learning. Similarly, Long Short-Term Memory (LSTM) has a strong inductive bias towards storing information over time. However, many real-world systems are governed by conservation laws, which lead to the redistribution of particular quantities {—} e.g.in physical and economical systems. Our novel Mass-Conserving LSTM (MC-LSTM) adheres to these conservation laws by extending the inductive bias of LSTM to model the redistribution of those stored quantities. MC-LSTMs set a new state-of-the-art for neural arithmetic units at learning arithmetic operations, such as addition tasks,which have a strong conservation law, as the sum is constant over time. Further, MC-LSTM is applied to traffic forecasting, modeling a pendulum, and a large benchmark dataset in hydrology, where it sets a new state-of-the-art for predicting peak flows. In the hydrology example, we show that MC-LSTM states correlate with real world processes and are therefore interpretable.
****

Learning Curves for Analysis of Deep Networks
Derek Hoiem, Tanmay Gupta, Zhizhong Li, Michal Shlapentokh-Rothman
Learning curves model a classifier’s test error as a function of the number of training samples. Prior works show that learning curves can be used to select model parameters and extrapolate performance. We investigate how to use learning curves to evaluate design choices, such as pretraining, architecture, and data augmentation. We propose a method to robustly estimate learning curves, abstract their parameters into error and data-reliance, and evaluate the effectiveness of different parameterizations. Our experiments exemplify use of learning curves for analysis and yield several interesting observations.
****

Equivariant Learning of Stochastic Fields: Gaussian Processes and Steerable Conditional Neural Processes
Peter Holderrieth, Michael J Hutchinson, Yee Whye Teh
Motivated by objects such as electric fields or fluid streams, we study the problem of learning stochastic fields, i.e. stochastic processes whose samples are fields like those occurring in physics and engineering. Considering general transformations such as rotations and reflections, we show that spatial invariance of stochastic fields requires an inference model to be equivariant. Leveraging recent advances from the equivariance literature, we study equivariance in two clas

ses of models. Firstly, we fully characterise equivariant Gaussian processes. Secondly, we introduce Steerable Conditional Neural Processes (SteerCNPs), a new, fully equivariant member of the Neural Process family. In experiments with Gaussian process vector fields, images, and real-world weather data, we observe that SteerCNPs significantly improve the performance of previous models and equivariance leads to improvements in transfer learning tasks.
****
Latent Programmer: Discrete Latent Codes for Program Synthesis
Joey Hong, David Dohan, Rishabh Singh, Charles Sutton, Manzil Zaheer
A key problem in program synthesis is searching over the large space of possible programs. Human programmers might decide the high-level structure of the desired program before thinking about the details; motivated by this intuition, we consider two-level search for program synthesis, in which the synthesizer first generates a plan, a sequence of symbols that describes the desired program at a high level, before generating the program. We propose to learn representations of programs that can act as plans to organize such a two-level search. Discrete latent codes are appealing for this purpose, and can be learned by applying recent work on discrete autoencoders. Based on these insights, we introduce the Latent Programmer (LP), a program synthesis method that first predicts a discrete latent code from input/output examples, and then generates the program in the target language. We evaluate the LP on two domains, demonstrating that it yields an improvement in accuracy, especially on longer programs for which search is most difficult.
****
Chebyshev Polynomial Codes: Task Entanglement-based Coding for Distributed Matrix Multiplication
Sangwoo Hong, Heecheol Yang, Youngseok Yoon, Taehyun Cho, Jungwoo Lee
Distributed computing has been a prominent solution to efficiently process massive datasets in parallel. However, the existence of stragglers is one of the major concerns that slows down the overall speed of distributed computing. To deal with this problem, we consider a distributed matrix multiplication scenario where a master assigns multiple tasks to each worker to exploit stragglers’ computing ability (which is typically wasted in conventional distributed computing). We propose Chebyshev polynomial codes, which can achieve order-wise improvement in encoding complexity at the master and communication load in distributed matrix multiplication using task entanglement. The key idea of task entanglement is to reduce the number of encoded matrices for multiple tasks assigned to each worker by intertwining encoded matrices. We experimentally demonstrate that, in cloud environments, Chebyshev polynomial codes can provide significant reduction in overall processing time in distributed computing for matrix multiplication, which is a key computational component in modern deep learning.
****
Federated Learning of User Verification Models Without Sharing Embeddings
Hossein Hosseini, Hyunsin Park, Sungrack Yun, Christos Louizos, Joseph Soriaga, Max Welling
We consider the problem of training User Verification (UV) models in federated setup, where each user has access to the data of only one class and user embeddings cannot be shared with the server or other users. To address this problem, we propose Federated User Verification (FedUV), a framework in which users jointly learn a set of vectors and maximize the correlation of their instance embeddings with a secret linear combination of those vectors. We show that choosing the linear combinations from the codewords of an error-correcting code allows users to collaboratively train the model without revealing their embedding vectors. We present the experimental results for user verification with voice, face, and handwriting data and show that FedUV is on par with existing approaches, while not sharing the embeddings with other users or the server.
****
The Limits of Min-Max Optimization Algorithms: Convergence to Spurious Non-Critical Sets
Ya-Ping Hsieh, Panayotis Mertikopoulos, Volkan Cevher

Compared to minimization, the min-max optimization in machine learning applicati
ons is considerably more convoluted because of the existence of cycles and simil
ar phenomena. Such oscillatory behaviors are well-understood in the convex-conca
ve regime, and many algorithms are known to overcome them. In this paper, we go
beyond this basic setting and characterize the convergence properties of many po
pular methods in solving non-convex/non-concave problems. In particular, we show
 that a wide class of state-of-the-art schemes and heuristics may converge with
arbitrarily high probability to attractors that are in no way min-max optimal or
 even stationary. Our work thus points out a potential pitfall among many existi
ng theoretical frameworks, and we corroborate our theoretical claims by explicit
ly showcasing spurious attractors in simple two-dimensional problems.
****
Near-Optimal Representation Learning for Linear Bandits and Linear RL
Jiachen Hu, Xiaoyu Chen, Chi Jin, Lihong Li, Liwei Wang
This paper studies representation learning for multi-task linear bandits and mul
ti-task episodic RL with linear value function approximation. We first consider
the setting where we play M linear bandits with dimension d concurrently, an
d these bandits share a common k-dimensional linear representation so that $k\
ll d$ and $k \ll M$. We propose a sample-efficient algorithm, MTLR-OFUL, which l
everages the shared representation to achieve $\tilde{O}(M\sqrt{dkT} + d\sqrt{kM
T})$ regret, with T being the number of total steps. Our regret significantly
 improves upon the baseline $\tilde{O}(Md\sqrt{T})$ achieved by solving each tas
k independently. We further develop a lower bound that shows our regret is near-
optimal when $d > M$. Furthermore, we extend the algorithm and analysis to multi
-task episodic RL with linear value function approximation under low inherent Be
llman error (Zanette et al., 2020a). To the best of our knowledge, this is the f
irst theoretical result that characterize the benefits of multi-task representat
ion learning for exploration in RL with function approximation.
****
On the Random Conjugate Kernel and Neural Tangent Kernel
Zhengmian Hu, Heng Huang
We investigate the distributions of Conjugate Kernel (CK) and Neural Tangent Ker
nel (NTK) for ReLU networks with random initialization. We derive the precise di
stributions and moments of the diagonal elements of these kernels. For a feedfor
ward network, these values converge in law to a log-normal distribution when the
 network depth d and width n simultaneously tend to infinity and the varianc
e of log diagonal elements is proportional to ${d}/{n}$. For the residual networ
k, in the limit that number of branches m increases to infinity and the width
n remains fixed, the diagonal elements of Conjugate Kernel converge in law to
a log-normal distribution where the variance of log value is proportional to ${1
}/{n}$, and the diagonal elements of NTK converge in law to a log-normal distrib
uted variable times the conjugate kernel of one feedforward network. Our new the
oretical analysis results suggest that residual network remains trainable in the
 limit of infinite branches and fixed network width. The numerical experiments a
re conducted and all results validate the soundness of our theoretical analysis.
****
Off-Belief Learning
Hengyuan Hu, Adam Lerer, Brandon Cui, Luis Pineda, Noam Brown, Jakob Foerster
The standard problem setting in Dec-POMDPs is self-play, where the goal is to fi
nd a set of policies that play optimally together. Policies learned through self
-play may adopt arbitrary conventions and implicitly rely on multi-step reasonin
g based on fragile assumptions about other agents’ actions and thus fail when pa
ired with humans or independently trained agents at test time. To address this,
we present off-belief learning (OBL). At each timestep OBL agents follow a polic
y π_1 that is optimized assuming past actions were taken by a given, fixed p
olicy (π_0), but assuming that future actions will be taken by π_1. When
 π_0 is uniform random, OBL converges to an optimal policy that does not rel
y on inferences based on other agents’ behavior (an optimal grounded policy). OB
L can be iterated in a hierarchy, where the optimal policy from one level become
s the input to the next, thereby introducing multi-level cognitive reasoning in

a controlled manner. Unlike existing approaches, which may converge to any equil ibrium policy, OBL converges to a unique policy, making it suitable for zero-sho t coordination (ZSC). OBL can be scaled to high-dimensional settings with a fict itious transition mechanism and shows strong performance in both a toy-setting a nd the benchmark human-AI & ZSC problem Hanabi.
****
Generalizable Episodic Memory for Deep Reinforcement Learning
Hao Hu, Jianing Ye, Guangxiang Zhu, Zhizhou Ren, Chongjie Zhang
Episodic memory-based methods can rapidly latch onto past successful strategies by a non-parametric memory and improve sample efficiency of traditional reinforc ement learning. However, little effort is put into the continuous domain, where a state is never visited twice, and previous episodic methods fail to efficientl y aggregate experience across trajectories. To address this problem, we propose Generalizable Episodic Memory (GEM), which effectively organizes the state-actio n values of episodic memory in a generalizable manner and supports implicit plan ning on memorized trajectories. GEM utilizes a double estimator to reduce the ov erestimation bias induced by value propagation in the planning process. Empirica l evaluation shows that our method significantly outperforms existing trajectory -based methods on various MuJoCo continuous control tasks. To further show the g eneral applicability, we evaluate our method on Atari games with discrete action space, which also shows a significant improvement over baseline algorithms.
****
A Scalable Deterministic Global Optimization Algorithm for Clustering Problems
Kaixun Hua, Mingfei Shi, Yankai Cao
The minimum sum-of-squares clustering (MSSC) task, which can be treated as a Mix ed Integer Second Order Cone Programming (MISOCP) problem, is rarely investigate d in the literature through deterministic optimization to find its global optima l value. In this paper, we modelled the MSSC task as a two-stage optimization pr oblem and proposed a tailed reduced-space branch and bound (BB) algorithm. We de signed several approaches to construct lower and upper bounds at each node in th e BB scheme, including a scenario grouping based Lagrangian decomposition approa ch. One key advantage of this reduced-space algorithm is that it only needs to p erform branching on the centers of clusters to guarantee convergence, and the si ze of centers is independent of the number of data samples. Moreover, the lower bounds can be computed by solving small-scale sample subproblems, and upper boun ds can be obtained trivially. These two properties enable our algorithm easy to be paralleled and can be scalable to the dataset with up to 200,000 samples for finding a global ϵ-optimal solution of the MSSC task. We performed nume rical experiments on both synthetic and real-world datasets and compared our pro posed algorithms with the off-the-shelf global optimal solvers and classical loc al optimal algorithms. The results reveal a strong performance and scalability o f our algorithm.
****
On Recovering from Modeling Errors Using Testing Bayesian Networks
Haiying Huang, Adnan Darwiche
We consider the problem of supervised learning with Bayesian Networks when the u sed dependency structure is incomplete due to missing edges or missing variable states. These modeling errors induce independence constraints on the learned mod el that may not hold in the true, data-generating distribution. We provide a uni fied treatment of these modeling errors as instances of state-space abstractions . We then identify a class of Bayesian Networks and queries which allow one to f ully recover from such modeling errors if one can choose Conditional Probability Tables (CPTs) dynamically based on evidence. We show theoretically that the rec ently proposed Testing Bayesian Networks (TBNs), which can be trained by compili ng them into Testing Arithmetic Circuits (TACs), provide a promising construct f or emulating this CPT selection mechanism. Finally, we present empirical results that illustrate the promise of TBNs as a tool for recovering from certain model ing errors in the context of supervised learning.
****
A Novel Sequential Coreset Method for Gradient Descent Algorithms

Jiawei Huang, Ruomin Huang, Wenjie Liu, Nikolaos Freris, Hu Ding
A wide range of optimization problems arising in machine learning can be solved by gradient descent algorithms, and a central question in this area is how to efficiently compress a large-scale dataset so as to reduce the computational complexity. Coreset is a popular data compression technique that has been extensively studied before. However, most of existing coreset methods are problem-dependent and cannot be used as a general tool for a broader range of applications. A key obstacle is that they often rely on the pseudo-dimension and total sensitivity bound that can be very high or hard to obtain. In this paper, based on the “locality” property of gradient descent algorithms, we propose a new framework, termed “sequential coreset”, which effectively avoids these obstacles. Moreover, our method is particularly suitable for sparse optimization whence the coreset size can be further reduced to be only poly-logarithmically dependent on the dimension. In practice, the experimental results suggest that our method can save a large amount of running time compared with the baseline algorithms.
****
FL-NTK: A Neural Tangent Kernel-based Framework for Federated Learning Analysis
Baihe Huang, Xiaoxiao Li, Zhao Song, Xin Yang
Federated Learning (FL) is an emerging learning scheme that allows different distributed clients to train deep neural networks together without data sharing. Neural networks have become popular due to their unprecedented success. To the best of our knowledge, the theoretical guarantees of FL concerning neural networks with explicit forms and multi-step updates are unexplored. Nevertheless, training analysis of neural networks in FL is non-trivial for two reasons: first, the objective loss function we are optimizing is non-smooth and non-convex, and second, we are even not updating in the gradient direction. Existing convergence results for gradient descent-based methods heavily rely on the fact that the gradient direction is used for updating. The current paper presents a new class of convergence analysis for FL, Federated Neural Tangent Kernel (FL-NTK), which corresponds to overparamterized ReLU neural networks trained by gradient descent in FL and is inspired by the analysis in Neural Tangent Kernel (NTK). Theoretically, FL-NTK converges to a global-optimal solution at a linear rate with properly tuned learning parameters. Furthermore, with proper distributional assumptions, FL-NTK can also achieve good generalization. The proposed theoretical analysis scheme can be generalized to more complex neural networks.
****
STRODE: Stochastic Boundary Ordinary Differential Equation
Hengguan Huang, Hongfu Liu, Hao Wang, Chang Xiao, Ye Wang
Perception of time from sequentially acquired sensory inputs is rooted in everyday behaviors of individual organisms. Yet, most algorithms for time-series modeling fail to learn dynamics of random event timings directly from visual or audio inputs, requiring timing annotations during training that are usually unavailable for real-world applications. For instance, neuroscience perspectives on postdiction imply that there exist variable temporal ranges within which the incoming sensory inputs can affect the earlier perception, but such temporal ranges are mostly unannotated for real applications such as automatic speech recognition (ASR). In this paper, we present a probabilistic ordinary differential equation (ODE), called STochastic boundaRy ODE (STRODE), that learns both the timings and the dynamics of time series data without requiring any timing annotations during training. STRODE allows the usage of differential equations to sample from the posterior point processes, efficiently and analytically. We further provide theoretical guarantees on the learning of STRODE. Our empirical results show that our approach successfully infers event timings of time series data. Our method achieves competitive or superior performances compared to existing state-of-the-art methods for both synthetic and real-world datasets.
****
A Riemannian Block Coordinate Descent Method for Computing the Projection Robust Wasserstein Distance
Minhui Huang, Shiqian Ma, Lifeng Lai
The Wasserstein distance has become increasingly important in machine learning a

nd deep learning. Despite its popularity, the Wasserstein distance is hard to ap proximate because of the curse of dimensionality. A recently proposed approach t o alleviate the curse of dimensionality is to project the sampled data from the high dimensional probability distribution onto a lower-dimensional subspace, and then compute the Wasserstein distance between the projected data. However, this approach requires to solve a max-min problem over the Stiefel manifold, which i s very challenging in practice. In this paper, we propose a Riemannian block coo rdinate descent (RBCD) method to solve this problem, which is based on a novel r eformulation of the regularized max-min problem over the Stiefel manifold. We sh ow that the complexity of arithmetic operations for RBCD to obtain an ϵ -stationary point is $O(\epsilon^{-3})$, which is significantly better than the complexity of existing methods. Numerical results on both synthetic and real dat asets demonstrate that our method is more efficient than existing methods, espec ially when the number of sampled data is very large.
****

Projection Robust Wasserstein Barycenters

Minhui Huang, Shiqian Ma, Lifeng Lai

Collecting and aggregating information from several probability measures or hist ograms is a fundamental task in machine learning. One of the popular solution me thods for this task is to compute the barycenter of the probability measures und er the Wasserstein metric. However, approximating the Wasserstein barycenter is numerically challenging because of the curse of dimensionality. This paper propo ses the projection robust Wasserstein barycenter (PRWB) that has the potential t o mitigate the curse of dimensionality, and a relaxed PRWB (RPRWB) model that is computationally more tractable. By combining the iterative Bregman projection a lgorithm and Riemannian optimization, we propose two algorithms for computing th e RPRWB, which is a max-min problem over the Stiefel manifold. The complexity of arithmetic operations of the proposed algorithms for obtaining an ϵ-st ationary solution is analyzed. We incorporate the RPRWB into a discrete distribu tion clustering algorithm, and the numerical results on real text datasets confi rm that our RPRWB model helps improve the clustering performance significantly.
****

Accurate Post Training Quantization With Small Calibration Sets

Itay Hubara, Yury Nahshan, Yair Hanani, Ron Banner, Daniel Soudry

Lately, post-training quantization methods have gained considerable attention, a s they are simple to use, and require only a small unlabeled calibration set. Th is small dataset cannot be used to fine-tune the model without significant over-fitting. Instead, these methods only use the calibration set to set the activati ons' dynamic ranges. However, such methods always resulted in significant accura cy degradation, when used below 8-bits (except on small datasets). Here we aim t o break the 8-bit barrier. To this end, we minimize the quantization errors of e ach layer or block separately by optimizing its parameters over the calibration set. We empirically demonstrate that this approach is: (1) much less susceptible to over-fitting than the standard fine-tuning approaches, and can be used even on a very small calibration set; and (2) more powerful than previous methods, wh ich only set the activations' dynamic ranges. We suggest two flavors for our met hod, parallel and sequential aim for a fixed and flexible bit-width allocation. For the latter, we demonstrate how to optimally allocate the bit-widths for each layer, while constraining accuracy degradation or model compression by proposin g a novel integer programming formulation. Finally, we suggest model global stat istics tuning, to correct biases introduced during quantization. Together, these methods yield state-of-the-art results for both vision and text models. For ins tance, on ResNet50, we obtain less than 1% accuracy degradation — with 4-bit wei ghts and activations in all layers, but first and last. The suggested methods ar e two orders of magnitude faster than the traditional Quantize Aware Training ap proach used for lower than 8-bit quantization. We open-sourced our code \textit{ https://github.com/papers-submission/CalibTIP}.
****

Learning and Planning in Complex Action Spaces

Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Mohammadamin Barekatain

, Simon Schmitt, David Silver
Many important real-world problems have action spaces that are high-dimensional, continuous or both, making full enumeration of all possible actions infeasible. Instead, only small subsets of actions can be sampled for the purpose of policy evaluation and improvement. In this paper, we propose a general framework to reason in a principled way about policy evaluation and improvement over such sampled action subsets. This sample-based policy iteration framework can in principle be applied to any reinforcement learning algorithm based upon policy iteration. Concretely, we propose Sampled MuZero, an extension of the MuZero algorithm that is able to learn in domains with arbitrarily complex action spaces by planning over sampled actions. We demonstrate this approach on the classical board game of Go and on two continuous control benchmark domains: DeepMind Control Suite and Real-World RL Suite.
****
Generative Adversarial Transformers
Drew A Hudson, Larry Zitnick
We introduce the GANsformer, a novel and efficient type of transformer, and explore it for the task of visual generative modeling. The network employs a bipartite structure that enables long-range interactions across the image, while maintaining computation of linear efficiency, that can readily scale to high-resolution synthesis. It iteratively propagates information from a set of latent variables to the evolving visual features and vice versa, to support the refinement of each in light of the other and encourage the emergence of compositional representations of objects and scenes. In contrast to the classic transformer architecture, it utilizes multiplicative integration that allows flexible region-based modulation, and can thus be seen as a generalization of the successful StyleGAN network. We demonstrate the model’s strength and robustness through a careful evaluation over a range of datasets, from simulated multi-object environments to rich real-world indoor and outdoor scenes, showing it achieves state-of-the-art results in terms of image quality and diversity, while enjoying fast learning and better data-efficiency. Further qualitative and quantitative experiments offer us an insight into the model’s inner workings, revealing improved interpretability and stronger disentanglement, and illustrating the benefits and efficacy of our approach. An implementation of the model is available at https://github.com/dorarad/gansformer.
****
Neural Pharmacodynamic State Space Modeling
Zeshan M Hussain, Rahul G. Krishnan, David Sontag
Modeling the time-series of high-dimensional, longitudinal data is important for predicting patient disease progression. However, existing neural network based approaches that learn representations of patient state, while very flexible, are susceptible to overfitting. We propose a deep generative model that makes use of a novel attention-based neural architecture inspired by the physics of how treatments affect disease state. The result is a scalable and accurate model of high-dimensional patient biomarkers as they vary over time. Our proposed model yields significant improvements in generalization and, on real-world clinical data, provides interpretable insights into the dynamics of cancer progression.
****
Hyperparameter Selection for Imitation Learning
Léonard Hussenot, Marcin Andrychowicz, Damien Vincent, Robert Dadashi, Anton Raichuk, Sabela Ramos, Nikola Momchev, Sertan Girgin, Raphael Marinier, Lukasz Stafiniak, Manu Orsini, Olivier Bachem, Matthieu Geist, Olivier Pietquin
We address the issue of tuning hyperparameters (HPs) for imitation learning algorithms in the context of continuous-control, when the underlying reward function of the demonstrating expert cannot be observed at any time. The vast literature in imitation learning mostly considers this reward function to be available for HP selection, but this is not a realistic setting. Indeed, would this reward function be available, it could then directly be used for policy training and imitation would not be necessary. To tackle this mostly ignored problem, we propose a number of possible proxies to the external reward. We evaluate them in an exte

nsive empirical study (more than 10’000 agents across 9 environments) and make practical recommendations for selecting HPs. Our results show that while imitation learning algorithms are sensitive to HP choices, it is often possible to select good enough HPs through a proxy to the reward function.
****
Pareto GAN: Extending the Representational Power of GANs to Heavy-Tailed Distributions
Todd Huster, Jeremy Cohen, Zinan Lin, Kevin Chan, Charles Kamhoua, Nandi O. Leslie, Cho-Yu Jason Chiang, Vyas Sekar
Generative adversarial networks (GANs) are often billed as "universal distribution learners", but precisely what distributions they can represent and learn is still an open question. Heavy-tailed distributions are prevalent in many different domains such as financial risk-assessment, physics, and epidemiology. We observe that existing GAN architectures do a poor job of matching the asymptotic behavior of heavy-tailed distributions, a problem that we show stems from their construction. Additionally, common loss functions produce unstable or near-zero gradients when faced with the infinite moments and large distances between outlier points characteristic of heavy-tailed distributions. We address these problems with the Pareto GAN. A Pareto GAN leverages extreme value theory and the functional properties of neural networks to learn a distribution that matches the asymptotic behavior of the marginal distributions of the features. We identify issues with standard loss functions and propose the use of alternative metric spaces that enable stable and efficient learning. Finally, we evaluate our proposed approach on a variety of heavy-tailed datasets.
****
LieTransformer: Equivariant Self-Attention for Lie Groups
Michael J Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, Hyunjik Kim
Group equivariant neural networks are used as building blocks of group invariant neural networks, which have been shown to improve generalisation performance and data efficiency through principled parameter sharing. Such works have mostly focused on group equivariant convolutions, building on the result that group equivariant linear maps are necessarily convolutions. In this work, we extend the scope of the literature to self-attention, that is emerging as a prominent building block of deep learning models. We propose the LieTransformer, an architecture composed of LieSelfAttention layers that are equivariant to arbitrary Lie groups and their discrete subgroups. We demonstrate the generality of our approach by showing experimental results that are competitive to baseline methods on a wide range of tasks: shape counting on point clouds, molecular property regression and modelling particle trajectories under Hamiltonian dynamics.
****
Crowdsourcing via Annotator Co-occurrence Imputation and Provable Symmetric Nonnegative Matrix Factorization
Shahana Ibrahim, Xiao Fu
Unsupervised learning of the Dawid-Skene (D&S) model from noisy, incomplete and crowdsourced annotations has been a long-standing challenge, and is a critical step towards reliably labeling massive data. A recent work takes a coupled nonnegative matrix factorization (CNMF) perspective, and shows appealing features: It ensures the identifiability of the D&S model and enjoys low sample complexity, as only the estimates of the co-occurrences of annotator labels are involved. However, the identifiability holds only when certain somewhat restrictive conditions are met in the context of crowdsourcing. Optimizing the CNMF criterion is also costly—and convergence assurances are elusive. This work recasts the pairwise co-occurrence based D&S model learning problem as a symmetric NMF (SymNMF) problem—which offers enhanced identifiability relative to CNMF. In practice, the SymNMF model is often (largely) incomplete, due to the lack of co-labeled items by some annotators. Two lightweight algorithms are proposed for co-occurrence imputation. Then, a low-complexity shifted rectified linear unit (ReLU)-empowered SymNMF algorithm is proposed to identify the D&S model. Various performance characterizations (e.g., missing co-occurrence recoverability, stability, and convergence

) and evaluations are also presented.
****
Selecting Data Augmentation for Simulating Interventions
Maximilian Ilse, Jakub M Tomczak, Patrick Forré
Machine learning models trained with purely observational data and the principle of empirical risk minimization (Vapnik 1992) can fail to generalize to unseen domains. In this paper, we focus on the case where the problem arises through spurious correlation between the observed domains and the actual task labels. We find that many domain generalization methods do not explicitly take this spurious correlation into account. Instead, especially in more application-oriented research areas like medical imaging or robotics, data augmentation techniques that are based on heuristics are used to learn domain invariant features. To bridge the gap between theory and practice, we develop a causal perspective on the problem of domain generalization. We argue that causal concepts can be used to explain the success of data augmentation by describing how they can weaken the spurious correlation between the observed domains and the task labels. We demonstrate that data augmentation can serve as a tool for simulating interventional data. We use these theoretical insights to derive a simple algorithm that is able to select data augmentation techniques that will lead to better domain generalization.
****
Scalable Marginal Likelihood Estimation for Model Selection in Deep Learning
Alexander Immer, Matthias Bauer, Vincent Fortuin, Gunnar Rätsch, Khan Mohammad Emtiyaz
Marginal-likelihood based model-selection, even though promising, is rarely used in deep learning due to estimation difficulties. Instead, most approaches rely on validation data, which may not be readily available. In this work, we present a scalable marginal-likelihood estimation method to select both hyperparameters and network architectures, based on the training data alone. Some hyperparameters can be estimated online during training, simplifying the procedure. Our marginal-likelihood estimate is based on Laplace’s method and Gauss-Newton approximations to the Hessian, and it outperforms cross-validation and manual tuning on standard regression and image classification datasets, especially in terms of calibration and out-of-distribution detection. Our work shows that marginal likelihoods can improve generalization and be useful when validation data is unavailable (e.g., in nonstationary settings).
****
Active Learning for Distributionally Robust Level-Set Estimation
Yu Inatsu, Shogo Iwazaki, Ichiro Takeuchi
Many cases exist in which a black-box function f with high evaluation cost depends on two types of variables $\bm x$ and $\bm w$, where $\bm x$ is a controllable \emph{design} variable and $\bm w$ are uncontrollable \emph{environmental} variables that have random variation following a certain distribution P. In such cases, an important task is to find the range of design variables $\bm x$ such that the function $f(\bm x, \bm w)$ has the desired properties by incorporating the random variation of the environmental variables $\bm w$. A natural measure of robustness is the probability that $f(\bm x, \bm w)$ exceeds a given threshold h, which is known as the \emph{probability threshold robustness} (PTR) measure in the literature on robust optimization. However, this robustness measure cannot be correctly evaluated when the distribution P is unknown. In this study, we addressed this problem by considering the \textit{distributionally robust PTR} (DRPTR) measure, which considers the worst-case PTR within given candidate distributions. Specifically, we studied the problem of efficiently identifying a reliable set H, which is defined as a region in which the DRPTR measure exceeds a certain desired probability α, which can be interpreted as a level set estimation (LSE) problem for DRPTR. We propose a theoretically grounded and computationally efficient active learning method for this problem. We show that the proposed method has theoretical guarantees on convergence and accuracy, and confirmed through numerical experiments that the proposed method outperforms existing methods.
****

Learning Randomly Perturbed Structured Predictors for Direct Loss Minimization
Hedda Cohen Indelman, Tamir Hazan
Direct loss minimization is a popular approach for learning predictors over structured label spaces. This approach is computationally appealing as it replaces integration with optimization and allows to propagate gradients in a deep net using loss-perturbed prediction. Recently, this technique was extended to generative models, by introducing a randomized predictor that samples a structure from a randomly perturbed score function. In this work, we interpolate between these techniques by learning the variance of randomized structured predictors as well as their mean, in order to balance between the learned score function and the randomized noise. We demonstrate empirically the effectiveness of learning this balance in structured discrete spaces.
****
Randomized Entity-wise Factorization for Multi-Agent Reinforcement Learning
Shariq Iqbal, Christian A Schroeder De Witt, Bei Peng, Wendelin Boehmer, Shimon Whiteson, Fei Sha
Multi-agent settings in the real world often involve tasks with varying types and quantities of agents and non-agent entities; however, common patterns of behavior often emerge among these agents/entities. Our method aims to leverage these commonalities by asking the question: “What is the expected utility of each agent when only considering a randomly selected sub-group of its observed entities?” By posing this counterfactual question, we can recognize state-action trajectories within sub-groups of entities that we may have encountered in another task and use what we learned in that task to inform our prediction in the current one. We then reconstruct a prediction of the full returns as a combination of factors considering these disjoint groups of entities and train this “randomly factorized" value function as an auxiliary objective for value-based multi-agent reinforcement learning. By doing so, our model can recognize and leverage similarities across tasks to improve learning efficiency in a multi-task setting. Our approach, Randomized Entity-wise Factorization for Imagined Learning (REFIL), outperforms all strong baselines by a significant margin in challenging multi-task StarCraft micromanagement settings.
****
Randomized Exploration in Reinforcement Learning with General Value Function Approximation
Haque Ishfaq, Qiwen Cui, Viet Nguyen, Alex Ayoub, Zhuoran Yang, Zhaoran Wang, Doina Precup, Lin Yang
We propose a model-free reinforcement learning algorithm inspired by the popular randomized least squares value iteration (RLSVI) algorithm as well as the optimism principle. Unlike existing upper-confidence-bound (UCB) based approaches, which are often computationally intractable, our algorithm drives exploration by simply perturbing the training data with judiciously chosen i.i.d. scalar noises. To attain optimistic value function estimation without resorting to a UCB-style bonus, we introduce an optimistic reward sampling procedure. When the value functions can be represented by a function class $\mathcal{F}$, our algorithm achieves a worst-case regret bound of $\tilde{O}(\mathrm{poly}(d_EH)\sqrt{T})$ where T is the time elapsed, H is the planning horizon and d_E is the \emph{eluder dimension} of $\mathcal{F}$. In the linear setting, our algorithm reduces to LSVI-PHE, a variant of RLSVI, that enjoys an $\tilde{\mathcal{O}}(\sqrt{d^3H^3T})$ regret. We complement the theory with an empirical evaluation across known difficult exploration tasks.
****
Distributed Second Order Methods with Fast Rates and Compressed Communication
Rustem Islamov, Xun Qian, Peter Richtarik
We develop several new communication-efficient second-order methods for distributed optimization. Our first method, NEWTON-STAR, is a variant of Newton’s method from which it inherits its fast local quadratic rate. However, unlike Newton’s method, NEWTON-STAR enjoys the same per iteration communication cost as gradient descent. While this method is impractical as it relies on the use of certain unknown parameters characterizing the Hessian of the objective function at the opt

imum, it serves as the starting point which enables us to design practical varia nts thereof with strong theoretical guarantees. In particular, we design a stoch astic sparsification strategy for learning the unknown parameters in an iterativ e fashion in a communication efficient manner. Applying this strategy to NEWTON-STAR leads to our next method, NEWTON-LEARN, for which we prove local linear and superlinear rates independent of the condition number. When applicable, this me thod can have dramatically superior convergence behavior when compared to state-of-the-art methods. Finally, we develop a globalization strategy using cubic reg ularization which leads to our next method, CUBIC-NEWTON-LEARN, for which we pro ve global sublinear and linear convergence rates, and a fast superlinear rate. O ur results are supported with experimental results on real datasets, and show se veral orders of magnitude improvement on baseline and state-of-the-art methods i n terms of communication complexity.
****
What Are Bayesian Neural Network Posteriors Really Like?
Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, Andrew Gordon Gordon Wilson
The posterior over Bayesian neural network (BNN) parameters is extremely high-di mensional and non-convex. For computational reasons, researchers approximate thi s posterior using inexpensive mini-batch methods such as mean-field variational inference or stochastic-gradient Markov chain Monte Carlo (SGMCMC). To investiga te foundational questions in Bayesian deep learning, we instead use full batch H amiltonian Monte Carlo (HMC) on modern architectures. We show that (1) BNNs can achieve significant performance gains over standard training and deep ensembles; (2) a single long HMC chain can provide a comparable representation of the post erior to multiple shorter chains; (3) in contrast to recent studies, we find pos terior tempering is not needed for near-optimal performance, with little evidenc e for a “cold posterior” effect, which we show is largely an artifact of data au gmentation; (4) BMA performance is robust to the choice of prior scale, and rela tively similar for diagonal Gaussian, mixture of Gaussian, and logistic priors; (5) Bayesian neural networks show surprisingly poor generalization under domain shift; (6) while cheaper alternatives such as deep ensembles and SGMCMC can prov ide good generalization, their predictive distributions are distinct from HMC. N otably, deep ensemble predictive distributions are similarly close to HMC as sta ndard SGLD, and closer than standard variational inference.
****
How to Learn when Data Reacts to Your Model: Performative Gradient Descent
Zachary Izzo, Lexing Ying, James Zou
Performative distribution shift captures the setting where the choice of which M L model is deployed changes the data distribution. For example, a bank which use s the number of open credit lines to determine a customer’s risk of default on a loan may induce customers to open more credit lines in order to improve their c hances of being approved. Because of the interactions between the model and data distribution, finding the optimal model parameters is challenging. Works in thi s area have focused on finding stable points, which can be far from optimal. Her e we introduce \emph{performative gradient descent} (PerfGD), an algorithm for c omputing performatively optimal points. Under regularity assumptions on the perf ormative loss, PerfGD is the first algorithm which provably converges to an opti mal point. PerfGD explicitly captures how changes in the model affects the data distribution and is simple to use. We support our findings with theory and exper iments.
****
Perceiver: General Perception with Iterative Attention
Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, Joao C arreira
Biological systems understand the world by simultaneously processing high-dimens ional inputs from modalities as diverse as vision, audition, touch, propriocepti on, etc. The perception models used in deep learning on the other hand are desig ned for individual modalities, often relying on domain-specific assumptions such as the local grid structures exploited by virtually all existing vision models. These priors introduce helpful inductive biases, but also lock models to indivi

dual modalities. In this paper we introduce the Perceiver {–} a model that build
s upon Transformers and hence makes few architectural assumptions about the rela
tionship between its inputs, but that also scales to hundreds of thousands of in
puts, like ConvNets. The model leverages an asymmetric attention mechanism to it
eratively distill inputs into a tight latent bottleneck, allowing it to scale to
handle very large inputs. We show that this architecture is competitive with or
outperforms strong, specialized models on classification tasks across various m
odalities: images, point clouds, audio, video and video+audio. The Perceiver obt
ains performance comparable to ResNet-50 and ViT on ImageNet without 2D convolut
ions by directly attending to 50,000 pixels. It is also competitive in all modal
ities in AudioSet.
****
Imitation by Predicting Observations
Andrew Jaegle, Yury Sulsky, Arun Ahuja, Jake Bruce, Rob Fergus, Greg Wayne
Imitation learning enables agents to reuse and adapt the hard-won expertise of o
thers, offering a solution to several key challenges in learning behavior. Altho
ugh it is easy to observe behavior in the real-world, the underlying actions may
not be accessible. We present a new method for imitation solely from observatio
ns that achieves comparable performance to experts on challenging continuous con
trol tasks while also exhibiting robustness in the presence of observations unre
lated to the task. Our method, which we call FORM (for "Future Observation Rewar
d Model") is derived from an inverse RL objective and imitates using a model of
expert behavior learned by generative modelling of the expert’s observations, wi
thout needing ground truth actions. We show that FORM performs comparably to a s
trong baseline IRL method (GAIL) on the DeepMind Control Suite benchmark, while
outperforming GAIL in the presence of task-irrelevant features.
****
Local Correlation Clustering with Asymmetric Classification Errors
Jafar Jafarov, Sanchit Kalhan, Konstantin Makarychev, Yury Makarychev
In the Correlation Clustering problem, we are given a complete weighted graph $G
$ with its edges labeled as “similar" and “dissimilar" by a noisy binary classif
ier. For a clustering $\mathcal{C}$ of graph G, a similar edge is in disagreem
ent with $\mathcal{C}$, if its endpoints belong to distinct clusters; and a diss
imilar edge is in disagreement with $\mathcal{C}$ if its endpoints belong to the
same cluster. The disagreements vector, $\disagree$, is a vector indexed by the
vertices of G such that the v-th coordinate $\disagree_v$ equals the weight
of all disagreeing edges incident on v. The goal is to produce a clustering t
hat minimizes the ℓ_p norm of the disagreements vector for $p\geq 1$. We st
udy the ℓ_p objective in Correlation Clustering under the following assumpt
ion: Every similar edge has weight in $[\alpha\mathbf{w},\mathbf{w}]$ and every
dissimilar edge has weight at least $\alpha\mathbf{w}$ (where $\alpha \leq 1$ an
d $\mathbf{w}>0$ is a scaling parameter). We give an $O\left((\nicefrac{1}{\alph
a})^{\nicefrac{1}{2}-\nicefrac{1}{2p}}\cdot \log\nicefrac{1}{\alpha}\right)$ app
roximation algorithm for this problem. Furthermore, we show an almost matching c
onvex programming integrality gap.
****
Alternative Microfoundations for Strategic Classification
Meena Jagadeesan, Celestine Mendler-Dünner, Moritz Hardt
When reasoning about strategic behavior in a machine learning context it is temp
ting to combine standard microfoundations of rational agents with the statistica
l decision theory underlying classification. In this work, we argue that a direc
t combination of these ingredients leads to brittle solution concepts of limited
descriptive and prescriptive value. First, we show that rational agents with pe
rfect information produce discontinuities in the aggregate response to a decisio
n rule that we often do not observe empirically. Second, when any positive fract
ion of agents is not perfectly strategic, desirable stable points—where the clas
sifier is optimal for the data it entails—no longer exist. Third, optimal decisi
on rules under standard microfoundations maximize a measure of negative external
ity known as social burden within a broad class of assumptions about agent behav
ior. Recognizing these limitations we explore alternatives to standard microfoun

dations for binary classification. We describe desiderata that help navigate the space of possible assumptions about agent responses, and we then propose the noisy response model. Inspired by smoothed analysis and empirical observations, noisy response incorporates imperfection in the agent responses, which we show mitigates the limitations of standard microfoundations. Our model retains analytical tractability, leads to more robust insights about stable points, and imposes a lower social burden at optimality.
****
Robust Density Estimation from Batches: The Best Things in Life are (Nearly) Free
Ayush Jain, Alon Orlitsky
In many applications data are collected in batches, some potentially biased, corrupt, or even adversarial. Learning algorithms for this setting have therefore garnered considerable recent attention. In particular, a sequence of works has shown that all approximately piecewise polynomial distributions—and in particular all Gaussian, Gaussian-mixture, log-concave, low-modal, and monotone-hazard distributions—can be learned robustly in polynomial time. However, these results left open the question, stated explicitly in \cite{chen2020learning}, about the best possible sample complexity of such algorithms. We answer this question, showing that, perhaps surprisingly, up to logarithmic factors, the optimal sample complexity is the same as for genuine, non-adversarial, data! To establish the result, we reduce robust learning of approximately piecewise polynomial distributions to robust learning of the probability of all subsets of size at most k of a larger discrete domain, and learn these probabilities in optimal sample complexity linear in k regardless of the domain size. In simulations, the algorithm runs very quickly and estimates distributions to essentially the accuracy achieved when all adversarial batches are removed. The results also imply the first polynomial-time sample-optimal algorithm for robust interval-based classification based on batched data.
****
Instance-Optimal Compressed Sensing via Posterior Sampling
Ajil Jalal, Sushrut Karmalkar, Alex Dimakis, Eric Price
We characterize the measurement complexity of compressed sensing of signals drawn from a known prior distribution, even when the support of the prior is the entire space (rather than, say, sparse vectors). We show for Gaussian measurements and \emph{any} prior distribution on the signal, that the posterior sampling estimator achieves near-optimal recovery guarantees. Moreover, this result is robust to model mismatch, as long as the distribution estimate (e.g., from an invertible generative model) is close to the true distribution in Wasserstein distance. We implement the posterior sampling estimator for deep generative priors using Langevin dynamics, and empirically find that it produces accurate estimates with more diversity than MAP.
****
Fairness for Image Generation with Uncertain Sensitive Attributes
Ajil Jalal, Sushrut Karmalkar, Jessica Hoffmann, Alex Dimakis, Eric Price
This work tackles the issue of fairness in the context of generative procedures, such as image super-resolution, which entail different definitions from the standard classification setting. Moreover, while traditional group fairness definitions are typically defined with respect to specified protected groups – camouflaging the fact that these groupings are artificial and carry historical and political motivations – we emphasize that there are no ground truth identities. For instance, should South and East Asians be viewed as a single group or separate groups? Should we consider one race as a whole or further split by gender? Choosing which groups are valid and who belongs in them is an impossible dilemma and being “fair” with respect to Asians may require being “unfair” with respect to South Asians. This motivates the introduction of definitions that allow algorithms to be \emph{oblivious} to the relevant groupings. We define several intuitive notions of group fairness and study their incompatibilities and trade-offs. We show that the natural extension of demographic parity is strongly dependent on the grouping, and \emph{impossible} to achieve obliviously. On the other hand, the c

onceptually new definition we introduce, Conditional Proportional Representation , can be achieved obliviously through Posterior Sampling. Our experiments validate our theoretical results and achieve fair image reconstruction using state-of-the-art generative models.
****
Feature Clustering for Support Identification in Extreme Regions
Hamid Jalalzai, Rémi Leluc
Understanding the complex structure of multivariate extremes is a major challenge in various fields from portfolio monitoring and environmental risk management to insurance. In the framework of multivariate Extreme Value Theory, a common characterization of extremes' dependence structure is the angular measure. It is a suitable measure to work in extreme regions as it provides meaningful insights concerning the subregions where extremes tend to concentrate their mass. The present paper develops a novel optimization-based approach to assess the dependence structure of extremes. This support identification scheme rewrites as estimating clusters of features which best capture the support of extremes. The dimension reduction technique we provide is applied to statistical learning tasks such as feature clustering and anomaly detection. Numerical experiments provide strong empirical evidence of the relevance of our approach.
****
Improved Regret Bounds of Bilinear Bandits using Action Space Analysis
Kyoungseok Jang, Kwang-Sung Jun, Se-Young Yun, Wanmo Kang
We consider the bilinear bandit problem where the learner chooses a pair of arms , each from two different action spaces of dimension d_1 and d_2, respectively. The learner then receives a reward whose expectation is a bilinear function of the two chosen arms with an unknown matrix parameter $\Theta^*\in\mathbb{R}^{d_1 \times d_2}$ with rank r. Despite abundant applications such as drug discovery, the optimal regret rate is unknown for this problem, though it was conjectured to be $\tilde O(\sqrt{d_1d_2(d_1+d_2)r T})$ by Jun et al. (2019) where $\tilde O$ ignores polylogarithmic factors in T. In this paper, we make progress towards closing the gap between the upper and lower bound on the optimal regret. First, we reject the conjecture above by proposing algorithms that achieve the regret $\tilde O(\sqrt{d_1 d_2 (d_1+d_2) T})$ using the fact that the action space dimension $O(d_1+d_2)$ is significantly lower than the matrix parameter dimension $O(d_1 d_2)$. Second, we additionally devise an algorithm with better empirical performance than previous algorithms.
****
Inverse Decision Modeling: Learning Interpretable Representations of Behavior
Daniel Jarrett, Alihan Hüyük, Mihaela Van Der Schaar
Decision analysis deals with modeling and enhancing decision processes. A principal challenge in improving behavior is in obtaining a transparent *description* of existing behavior in the first place. In this paper, we develop an expressive, unifying perspective on *inverse decision modeling*: a framework for learning parameterized representations of sequential decision behavior. First, we formalize the *forward* problem (as a normative standard), subsuming common classes of control behavior. Second, we use this to formalize the *inverse* problem (as a descriptive model), generalizing existing work on imitation/reward learning—while opening up a much broader class of research problems in behavior representation. Finally, we instantiate this approach with an example (*inverse bounded rational control*), illustrating how this structure enables learning (interpretable) representations of (bounded) rationality—while naturally capturing intuitive notions of suboptimal actions, biased beliefs, and imperfect knowledge of environments.
****
Catastrophic Fisher Explosion: Early Phase Fisher Matrix Impacts Generalization
Stanislaw Jastrzebski, Devansh Arpit, Oliver Astrand, Giancarlo B Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, Krzysztof J Geras
The early phase of training a deep neural network has a dramatic effect on the local curvature of the loss function. For instance, using a small learning rate does not guarantee stable optimization because the optimization trajectory has a

tendency to steer towards regions of the loss surface with increasing local curvature. We ask whether this tendency is connected to the widely observed phenomenon that the choice of the learning rate strongly influences generalization. We first show that stochastic gradient descent (SGD) implicitly penalizes the trace of the Fisher Information Matrix (FIM), a measure of the local curvature, from the start of training. We argue it is an implicit regularizer in SGD by showing that explicitly penalizing the trace of the FIM can significantly improve generalization. We highlight that poor final generalization coincides with the trace of the FIM attaining a large value early in training, to which we refer as catastrophic Fisher explosion. Finally, to gain insight into the regularization effect of penalizing the trace of the FIM, we show that it limits memorization by reducing the learning speed of examples with noisy labels more than that of the examples with clean labels.
****
Policy Gradient Bayesian Robust Optimization for Imitation Learning
Zaynah Javed, Daniel S Brown, Satvik Sharma, Jerry Zhu, Ashwin Balakrishna, Marek Petrik, Anca Dragan, Ken Goldberg
The difficulty in specifying rewards for many real-world problems has led to an increased focus on learning rewards from human feedback, such as demonstrations. However, there are often many different reward functions that explain the human feedback, leaving agents with uncertainty over what the true reward function is. While most policy optimization approaches handle this uncertainty by optimizing for expected performance, many applications demand risk-averse behavior. We derive a novel policy gradient-style robust optimization approach, PG-BROIL, that optimizes a soft-robust objective that balances expected performance and risk. To the best of our knowledge, PG-BROIL is the first policy optimization algorithm robust to a distribution of reward hypotheses which can scale to continuous MDPs. Results suggest that PG-BROIL can produce a family of behaviors ranging from risk-neutral to risk-averse and outperforms state-of-the-art imitation learning algorithms when learning from ambiguous demonstrations by hedging against uncertainty, rather than seeking to uniquely identify the demonstrator’s reward function.
****
In-Database Regression in Input Sparsity Time
Rajesh Jayaram, Alireza Samadian, David Woodruff, Peng Ye
Sketching is a powerful dimensionality reduction technique for accelerating algorithms for data analysis. A crucial step in sketching methods is to compute a subspace embedding (SE) for a large matrix $A \in \mathbb{R}^{N \times d}$. SE’s are the primary tool for obtaining extremely efficient solutions for many linear-algebraic tasks, such as least squares regression and low rank approximation. Computing an SE often requires an explicit representation of A and running time proportional to the size of A. However, if $A= T_1 \Join T_2 \Join …\Join T_m$ is the result of a database join query on several smaller tables $T_i \in \mathbb{R}^{n_i \times d_i}$, then this running time can be prohibitive, as A itself can have as many as $O(n_1 n_2 \cdots n_m)$ rows. In this work, we design subspace embeddings for database joins which can be computed significantly faster than computing the join. For the case of a two table join $A = T_1 \Join T_2$ we give input-sparsity algorithms for computing subspace embeddings, with running time bounded by the number of non-zero entries in T_1,T_2. This results in input-sparsity time algorithms for high accuracy regression, significantly improving upon the running time of prior FAQ-based methods for regression. We extend our results to arbitrary joins for the ridge regression problem, also considerably improving the running time of prior methods. Empirically, we apply our method to real datasets and show that it is significantly faster than existing algorithms.
****
Parallel and Flexible Sampling from Autoregressive Models via Langevin Dynamics
Vivek Jayaram, John Thickstun
This paper introduces an alternative approach to sampling from autoregressive models. Autoregressive models are typically sampled sequentially, according to the transition dynamics defined by the model. Instead, we propose a sampling proced

ure that initializes a sequence with white noise and follows a Markov chain defined by Langevin dynamics on the global log-likelihood of the sequence. This approach parallelizes the sampling process and generalizes to conditional sampling. Using an autoregressive model as a Bayesian prior, we can steer the output of a generative model using a conditional likelihood or constraints. We apply these techniques to autoregressive models in the visual and audio domains, with competitive results for audio source separation, super-resolution, and inpainting.
****
Objective Bound Conditional Gaussian Process for Bayesian Optimization
Taewon Jeong, Heeyoung Kim
A Gaussian process is a standard surrogate model for an unknown objective function in Bayesian optimization. In this paper, we propose a new surrogate model, called the objective bound conditional Gaussian process (OBCGP), to condition a Gaussian process on a bound on the optimal function value. The bound is obtained and updated as the best observed value during the sequential optimization procedure. Unlike the standard Gaussian process, the OBCGP explicitly incorporates the existence of a point that improves the best known bound. We treat the location of such a point as a model parameter and estimate it jointly with other parameters by maximizing the likelihood using variational inference. Within the standard Bayesian optimization framework, the OBCGP can be combined with various acquisition functions to select the next query point. In particular, we derive cumulative regret bounds for the OBCGP combined with the upper confidence bound acquisition algorithm. Furthermore, the OBCGP can inherently incorporate a new type of prior knowledge, i.e., the bounds on the optimum, if it is available. The incorporation of this type of prior knowledge into a surrogate model has not been studied previously. We demonstrate the effectiveness of the OBCGP through its application to Bayesian optimization tasks, such as the sequential design of experiments and hyperparameter optimization in neural networks.
****
Quantifying Ignorance in Individual-Level Causal-Effect Estimates under Hidden Confounding
Andrew Jesson, Sören Mindermann, Yarin Gal, Uri Shalit
We study the problem of learning conditional average treatment effects (CATE) from high-dimensional, observational data with unobserved confounders. Unobserved confounders introduce ignorance—a level of unidentifiability—about an individual’s response to treatment by inducing bias in CATE estimates. We present a new parametric interval estimator suited for high-dimensional data, that estimates a range of possible CATE values when given a predefined bound on the level of hidden confounding. Further, previous interval estimators do not account for ignorance about the CATE associated with samples that may be underrepresented in the original study, or samples that violate the overlap assumption. Our interval estimator also incorporates model uncertainty so that practitioners can be made aware of such out-of-distribution data. We prove that our estimator converges to tight bounds on CATE when there may be unobserved confounding and assess it using semi-synthetic, high-dimensional datasets.
****
DeepReDuce: ReLU Reduction for Fast Private Inference
Nandan Kumar Jha, Zahra Ghodsi, Siddharth Garg, Brandon Reagen
The recent rise of privacy concerns has led researchers to devise methods for private neural inference—where inferences are made directly on encrypted data, never seeing inputs. The primary challenge facing private inference is that computing on encrypted data levies an impractically-high latency penalty, stemming mostly from non-linear operators like ReLU. Enabling practical and private inference requires new optimization methods that minimize network ReLU counts while preserving accuracy. This paper proposes DeepReDuce: a set of optimizations for the judicious removal of ReLUs to reduce private inference latency. The key insight is that not all ReLUs contribute equally to accuracy. We leverage this insight to drop, or remove, ReLUs from classic networks to significantly reduce inference latency and maintain high accuracy. Given a network architecture, DeepReDuce outputs a Pareto frontier of networks that tradeoff the number of ReLUs and accurac

y. Compared to the state-of-the-art for private inference DeepReDuce improves accuracy and reduces ReLU count by up to 3.5% (iso-ReLU count) and 3.5x (iso-accuracy), respectively.
****
Factor-analytic inverse regression for high-dimension, small-sample dimensionality reduction
Aditi Jha, Michael J. Morais, Jonathan W Pillow
Sufficient dimension reduction (SDR) methods are a family of supervised methods for dimensionality reduction that seek to reduce dimensionality while preserving information about a target variable of interest. However, existing SDR methods typically require more observations than the number of dimensions ($N > p$). To overcome this limitation, we propose Class-conditional Factor Analytic Dimensions (CFAD), a model-based dimensionality reduction method for high-dimensional, small-sample data. We show that CFAD substantially outperforms existing SDR methods in the small-sample regime, and can be extended to incorporate prior information such as smoothness in the projection axes. We demonstrate the effectiveness of CFAD with an application to functional magnetic resonance imaging (fMRI) measurements during visual object recognition and working memory tasks, where it outperforms existing SDR and a variety of other dimensionality-reduction methods.
****
Fast margin maximization via dual acceleration
Ziwei Ji, Nathan Srebro, Matus Telgarsky
We present and analyze a momentum-based gradient method for training linear classifiers with an exponentially-tailed loss (e.g., the exponential or logistic loss), which maximizes the classification margin on separable data at a rate of O(1/t^2). This contrasts with a rate of O(1/log(t)) for standard gradient descent, and O(1/t) for normalized gradient descent. The momentum-based method is derived via the convex dual of the maximum-margin problem, and specifically by applying Nesterov acceleration to this dual, which manages to result in a simple and intuitive method in the primal. This dual view can also be used to derive a stochastic variant, which performs adaptive non-uniform sampling via the dual variables.
****
Marginalized Stochastic Natural Gradients for Black-Box Variational Inference
Geng Ji, Debora Sujono, Erik B Sudderth
Black-box variational inference algorithms use stochastic sampling to analyze diverse statistical models, like those expressed in probabilistic programming languages, without model-specific derivations. While the popular score-function estimator computes unbiased gradient estimates, its variance is often unacceptably large, especially in models with discrete latent variables. We propose a stochastic natural gradient estimator that is as broadly applicable and unbiased, but improves efficiency by exploiting the curvature of the variational bound, and provably reduces variance by marginalizing discrete latent variables. Our marginalized stochastic natural gradients have intriguing connections to classic coordinate ascent variational inference, but allow parallel updates of variational parameters, and provide superior convergence guarantees relative to naive Monte Carlo approximations. We integrate our method with the probabilistic programming language Pyro and evaluate real-world models of documents, images, networks, and crowd-sourcing. Compared to score-function estimators, we require far fewer Monte Carlo samples and consistently convergence orders of magnitude faster.
****
Bilevel Optimization: Convergence Analysis and Enhanced Design
Kaiyi Ji, Junjie Yang, Yingbin Liang
Bilevel optimization has arisen as a powerful tool for many machine learning problems such as meta-learning, hyperparameter optimization, and reinforcement learning. In this paper, we investigate the nonconvex-strongly-convex bilevel optimization problem. For deterministic bilevel optimization, we provide a comprehensive convergence rate analysis for two popular algorithms respectively based on approximate implicit differentiation (AID) and iterative differentiation (ITD). For the AID-based method, we orderwisely improve the previous convergence rate ana

lysis due to a more practical parameter selection as well as a warm start strategy, and for the ITD-based method we establish the first theoretical convergence rate. Our analysis also provides a quantitative comparison between ITD and AID based approaches. For stochastic bilevel optimization, we propose a novel algorithm named stocBiO, which features a sample-efficient hypergradient estimator using efficient Jacobian- and Hessian-vector product computations. We provide the convergence rate guarantee for stocBiO, and show that stocBiO outperforms the best known computational complexities orderwisely with respect to the condition number κ and the target accuracy ϵ. We further validate our theoretical results and demonstrate the efficiency of bilevel optimization algorithms by the experiments on meta-learning and hyperparameter optimization.
****
Efficient Statistical Tests: A Neural Tangent Kernel Approach
Sheng Jia, Ehsan Nezhadarya, Yuhuai Wu, Jimmy Ba
For machine learning models to make reliable predictions in deployment, one needs to ensure the previously unknown test samples need to be sufficiently similar to the training data. The commonly used shift-invariant kernels do not have the compositionality and fail to capture invariances in high-dimensional data in computer vision. We propose a shift-invariant convolutional neural tangent kernel (SCNTK) based outlier detector and two-sample tests with maximum mean discrepancy (MMD) that is O(n) in the number of samples due to using the random feature approximation. On MNIST and CIFAR10 with various types of dataset shifts, we empirically show that statistical tests with such compositional kernels, inherited from infinitely wide neural networks, achieve higher detection accuracy than existing non-parametric methods. Our method also provides a competitive alternative to adapted kernel methods that require a training phase.
****
Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision
Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, Tom Duerig
Pre-trained representations are becoming crucial for many NLP and perception tasks. While representation learning in NLP has transitioned to training on raw text without human annotations, visual and vision-language representations still rely heavily on curated training datasets that are expensive or require expert knowledge. For vision applications, representations are mostly learned using datasets with explicit class labels such as ImageNet or OpenImages. For vision-language, popular datasets like Conceptual Captions, MSCOCO, or CLIP all involve a non-trivial data collection (and cleaning) process. This costly curation process limits the size of datasets and hence hinders the scaling of trained models. In this paper, we leverage a noisy dataset of over one billion image alt-text pairs, obtained without expensive filtering or post-processing steps in the Conceptual Captions dataset. A simple dual-encoder architecture learns to align visual and language representations of the image and text pairs using a contrastive loss. We show that the scale of our corpus can make up for its noise and leads to state-of-the-art representations even with such a simple learning scheme. Our visual representation achieves strong performance when transferred to classification tasks such as ImageNet and VTAB. The aligned visual and language representations enables zero-shot image classification and also set new state-of-the-art results on Flickr30K and MSCOCO image-text retrieval benchmarks, even when compared with more sophisticated cross-attention models. The representations also enable cross-modality search with complex text and text + image queries.
****
Multi-Dimensional Classification via Sparse Label Encoding
Bin-Bin Jia, Min-Ling Zhang
In multi-dimensional classification (MDC), there are multiple class variables in the output space with each of them corresponding to one heterogeneous class space. Due to the heterogeneity of class spaces, it is quite challenging to consider the dependencies among class variables when learning from MDC examples. In this paper, we propose a novel MDC approach named SLEM which learns the predictive

model in an encoded label space instead of the original heterogeneous one. Specifically, SLEM works in an encoding-training-decoding framework. In the encoding phase, each class vector is mapped into a real-valued one via three cascaded operations including pairwise grouping, one-hot conversion and sparse linear encoding. In the training phase, a multi-output regression model is learned within the encoded label space. In the decoding phase, the predicted class vector is obtained by adapting orthogonal matching pursuit over outputs of the learned multi-output regression model. Experimental results clearly validate the superiority of SLEM against state-of-the-art MDC approaches.
****
Self-Damaging Contrastive Learning
Ziyu Jiang, Tianlong Chen, Bobak J Mortazavi, Zhangyang Wang
The recent breakthrough achieved by contrastive learning accelerates the pace for deploying unsupervised training on real-world data applications. However, unlabeled data in reality is commonly imbalanced and shows a long-tail distribution, and it is unclear how robustly the latest contrastive learning methods could perform in the practical scenario. This paper proposes to explicitly tackle this challenge, via a principled framework called Self-Damaging Contrastive Learning (SDCLR), to automatically balance the representation learning without knowing the classes. Our main inspiration is drawn from the recent finding that deep models have difficult-to-memorize samples, and those may be exposed through network pruning. It is further natural to hypothesize that long-tail samples are also tougher for the model to learn well due to insufficient examples. Hence, the key innovation in SDCLR is to create a dynamic self-competitor model to contrast with the target model, which is a pruned version of the latter. During training, contrasting the two models will lead to adaptive online mining of the most easily forgotten samples for the current target model, and implicitly emphasize them more in the contrastive loss. Extensive experiments across multiple datasets and imbalance settings show that SDCLR significantly improves not only overall accuracies but also balancedness, in terms of linear evaluation on the full-shot and few-shot settings. Our code is available at https://github.com/VITA-Group/SDCLR.
****
Prioritized Level Replay
Minqi Jiang, Edward Grefenstette, Tim Rocktäschel
Environments with procedurally generated content serve as important benchmarks for testing systematic generalization in deep reinforcement learning. In this setting, each level is an algorithmically created environment instance with a unique configuration of its factors of variation. Training on a prespecified subset of levels allows for testing generalization to unseen levels. What can be learned from a level depends on the current policy, yet prior work defaults to uniform sampling of training levels independently of the policy. We introduce Prioritized Level Replay (PLR), a general framework for selectively sampling the next training level by prioritizing those with higher estimated learning potential when revisited in the future. We show TD-errors effectively estimate a level’s future learning potential and, when used to guide the sampling procedure, induce an emergent curriculum of increasingly difficult levels. By adapting the sampling of training levels, PLR significantly improves sample-efficiency and generalization on Procgen Benchmark—matching the previous state-of-the-art in test return—and readily combines with other methods. Combined with the previous leading method, PLR raises the state-of-the-art to over 76% improvement in test return relative to standard RL baselines.
****
Monotonic Robust Policy Optimization with Model Discrepancy
Yuankun Jiang, Chenglin Li, Wenrui Dai, Junni Zou, Hongkai Xiong
State-of-the-art deep reinforcement learning (DRL) algorithms tend to overfit due to the model discrepancy between source and target environments. Though applying domain randomization during training can improve the average performance by randomly generating a sufficient diversity of environments in simulator, the worst-case environment is still neglected without any performance guarantee. Since the average and worst-case performance are both important for generalization in R

L, in this paper, we propose a policy optimization approach for concurrently improving the policy’s performance in the average and worst-case environment. We theoretically derive a lower bound for the worst-case performance of a given policy by relating it to the expected performance. Guided by this lower bound, we formulate an optimization problem to jointly optimize the policy and sampling distribution, and prove that by iteratively solving it the worst-case performance is monotonically improved. We then develop a practical algorithm, named monotonic robust policy optimization (MRPO). Experimental evaluations in several robot control tasks demonstrate that MRPO can generally improve both the average and worst-case performance in the source environments for training, and facilitate in all cases the learned policy with a better generalization capability in some unseen testing environments.
****
Approximation Theory of Convolutional Architectures for Time Series Modelling
Haotian Jiang, Zhong Li, Qianxiao Li
We study the approximation properties of convolutional architectures applied to time series modelling, which can be formulated mathematically as a functional approximation problem. In the recurrent setting, recent results reveal an intricate connection between approximation efficiency and memory structures in the data generation process. In this paper, we derive parallel results for convolutional architectures, with WaveNet being a prime example. Our results reveal that in this new setting, approximation efficiency is not only characterised by memory, but also additional fine structures in the target relationship. This leads to a novel definition of spectrum-based regularity that measures the complexity of temporal relationships under the convolutional approximation scheme. These analyses provide a foundation to understand the differences between architectural choices for time series modelling and can give theoretically grounded guidance for practical applications.
****
Streaming and Distributed Algorithms for Robust Column Subset Selection
Shuli Jiang, Dennis Li, Irene Mengze Li, Arvind V Mahankali, David Woodruff
We give the first single-pass streaming algorithm for Column Subset Selection with respect to the entrywise ℓ_p-norm with $1 \leq p < 2$. We study the ℓ_p norm loss since it is often considered more robust to noise than the standard Frobenius norm. Given an input matrix $A \in \mathbb{R}^{d \times n}$ ($n \gg d$), our algorithm achieves a multiplicative $k^{\frac{1}{p} - \frac{1}{2}}\poly(\log nd)$-approximation to the error with respect to the \textit{best possible column subset} of size k. Furthermore, the space complexity of the streaming algorithm is optimal up to a logarithmic factor. Our streaming algorithm also extends naturally to a 1-round distributed protocol with nearly optimal communication cost. A key ingredient in our algorithms is a reduction to column subset selection in the $\ell_{p,2}$-norm, which corresponds to the p-norm of the vector of Euclidean norms of each of the columns of A. This enables us to leverage strong coreset constructions for the Euclidean norm, which previously had not been applied in this context. We also give the first provable guarantees for greedy column subset selection in the $\ell_{1, 2}$ norm, which can be used as an alternative, practical subroutine in our algorithms. Finally, we show that our algorithms give significant practical advantages on real-world data analysis tasks.
****
Single Pass Entrywise-Transformed Low Rank Approximation
Yifei Jiang, Yi Li, Yiming Sun, Jiaxin Wang, David Woodruff
In applications such as natural language processing or computer vision, one is given a large $n \times n$ matrix $A = (a_{i,j})$ and would like to compute a matrix decomposition, e.g., a low rank approximation, of a function $f(A) = (f(a_{i,j}))$ applied entrywise to A. A very important special case is the likelihood function $f\left(A \right) = \log{\left(\left| a_{ij}\right| +1\right)}$. A natural way to do this would be to simply apply f to each entry of A, and then compute the matrix decomposition, but this requires storing all of A as well as multiple passes over its entries. Recent work of Liang et al. shows how to find a rank-k factorization to $f(A)$ using only $n \cdot \poly(\eps^{-1}k\log

n)$ words of memory, with overall error $10\|f(A)-[f(A)]_k\|_F^2 + \poly(\epsilon/k) \|f(A)\|_{1,2}^2$, where $[f(A)]_k$ is the best rank-k approximation to $f(A)$ and $\|f(A)\|_{1,2}^2$ is the square of the sum of Euclidean lengths of rows of $f(A)$. Their algorithm uses 3 passes over the entries of A. The authors pose the open question of obtaining an algorithm with $n \cdot \poly(\eps^{-1}k\log n)$ words of memory using only a single pass over the entries of A. In this paper we resolve this open question, obtaining the first single-pass algorithm for this problem and for the same class of functions f studied by Liang et al. Moreover, our error is $\|f(A)-[f(A)]_k\|_F^2 + \poly(\epsilon/k) \|f(A)\|_F^2$, where $\|f(A)\|_F^2$ is the sum of squares of Euclidean lengths of rows of $f(A)$. Thus our error is significantly smaller, as it removes the factor of 10 and also $\|f(A)\|_F^2 \leq \|f(A)\|_{1,2}^2$.
****
The Emergence of Individuality
Jiechuan Jiang, Zongqing Lu
Individuality is essential in human society. It induces the division of labor and thus improves the efficiency and productivity. Similarly, it should also be a key to multi-agent cooperation. Inspired by that individuality is of being an individual separate from others, we propose a simple yet efficient method for the emergence of individuality (EOI) in multi-agent reinforcement learning (MARL). EOI learns a probabilistic classifier that predicts a probability distribution over agents given their observation and gives each agent an intrinsic reward of being correctly predicted by the classifier. The intrinsic reward encourages the agents to visit their own familiar observations, and learning the classifier by such observations makes the intrinsic reward signals stronger and in turn makes the agents more identifiable. To further enhance the intrinsic reward and promote the emergence of individuality, two regularizers are proposed to increase the discriminability of the classifier. We implement EOI on top of popular MARL algorithms. Empirically, we show that EOI outperforms existing methods in a variety of multi-agent cooperative scenarios.
****
Online Selection Problems against Constrained Adversary
Zhihao Jiang, Pinyan Lu, Zhihao Gavin Tang, Yuhao Zhang
Inspired by a recent line of work in online algorithms with predictions, we study the constrained adversary model that utilizes predictions from a different perspective. Prior works mostly focused on designing simultaneously robust and consistent algorithms, without making assumptions on the quality of the predictions. In contrary, our model assumes the adversarial instance is consistent with the predictions and aim to design algorithms that have best worst-case performance against all such instances. We revisit classical online selection problems under the constrained adversary model. For the single item selection problem, we design an optimal algorithm in the adversarial arrival model and an improved algorithm in the random arrival model (a.k.a., the secretary problem). For the online edge-weighted bipartite matching problem, we extend the classical Water-filling and Ranking algorithms and achieve improved competitive ratios.
****
Active Covering
Heinrich Jiang, Afshin Rostamizadeh
We analyze the problem of active covering, where the learner is given an unlabeled dataset and can sequentially label query examples. The objective is to label query all of the positive examples in the fewest number of total label queries. We show under standard non-parametric assumptions that a classical support estimator can be repurposed as an offline algorithm attaining an excess query cost of $\widetilde{\Theta}(n^{D/(D+1)})$ compared to the optimal learner, where n is the number of datapoints and D is the dimension. We then provide a simple active learning method that attains an improved excess query cost of $\widetilde{O}(n^{(D-1)/D})$. Furthermore, the proposed algorithms only require access to the positive labeled examples, which in certain settings provides additional computational and privacy benefits. Finally, we show that the active learning method consistently outperforms offline methods as well as a variety of baselines on a wi

de range of benchmark image-based datasets.
****
Emphatic Algorithms for Deep Reinforcement Learning
Ray Jiang, Tom Zahavy, Zhongwen Xu, Adam White, Matteo Hessel, Charles Blundell,
 Hado Van Hasselt
Off-policy learning allows us to learn about possible policies of behavior from
experience generated by a different behavior policy. Temporal difference (TD) le
arning algorithms can become unstable when combined with function approximation
and off-policy sampling—this is known as the “deadly triad”. Emphatic temporal d
ifference (ETD(λ)) algorithm ensures convergence in the linear case by a
ppropriately weighting the TD(λ) updates. In this paper, we extend the u
se of emphatic methods to deep reinforcement learning agents. We show that naive
ly adapting ETD(λ) to popular deep reinforcement learning algorithms, wh
ich use forward view multi-step returns, results in poor performance. We then de
rive new emphatic algorithms for use in the context of such algorithms, and we d
emonstrate that they provide noticeable benefits in small problems designed to h
ighlight the instability of TD methods. Finally, we observed improved performanc
e when applying these algorithms at scale on classic Atari games from the Arcade
 Learning Environment.
****
Characterizing Structural Regularities of Labeled Data in Overparameterized Mode
ls
Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, Michael C Mozer
Humans are accustomed to environments that contain both regularities and excepti
ons. For example, at most gas stations, one pays prior to pumping, but the occas
ional rural station does not accept payment in advance. Likewise, deep neural ne
tworks can generalize across instances that share common patterns or structures,
 yet have the capacity to memorize rare or irregular forms. We analyze how indiv
idual instances are treated by a model via a consistency score. The score charac
terizes the expected accuracy for a held-out instance given training sets of var
ying size sampled from the data distribution. We obtain empirical estimates of t
his score for individual instances in multiple data sets, and we show that the s
core identifies out-of-distribution and mislabeled examples at one end of the co
ntinuum and strongly regular examples at the other end. We identify computationa
lly inexpensive proxies to the consistency score using statistics collected duri
ng training. We apply the score toward understanding the dynamics of representat
ion learning and to filter outliers during training.
****
Optimal Streaming Algorithms for Multi-Armed Bandits
Tianyuan Jin, Keke Huang, Jing Tang, Xiaokui Xiao
This paper studies two variants of the best arm identification (BAI) problem und
er the streaming model, where we have a stream of n arms with reward distributio
ns supported on [0,1] with unknown means. The arms in the stream are arriving on
e by one, and the algorithm cannot access an arm unless it is stored in a limite
d size memory. We first study the streaming \epslion-topk-arms identification pr
oblem, which asks for k arms whose reward means are lower than that of the k-th
best arm by at most \epsilon with probability at least 1-\delta. For general \ep
silon \in (0,1), the existing solution for this problem assumes k = 1 and achiev
es the optimal sample complexity O(\frac{n}{\epsilon^2} \log \frac{1}{\delta}) u
sing O(\log^*(n)) memory and a single pass of the stream. We propose an algorith
m that works for any k and achieves the optimal sample complexity O(\frac{n}{\ep
silon^2} \log\frac{k}{\delta}) using a single-arm memory and a single pass of th
e stream. Second, we study the streaming BAI problem, where the objective is to
identify the arm with the maximum reward mean with at least 1-\delta probability
, using a single-arm memory and as few passes of the input stream as possible. W
e present a single-arm-memory algorithm that achieves a near instance-dependent
optimal sample complexity within O(\log \Delta_2^{-1}) passes, where \Delta_2 is
 the gap between the mean of the best arm and that of the second best arm.
****
Towards Tight Bounds on the Sample Complexity of Average-reward MDPs

Yujia Jin, Aaron Sidford
We prove new upper and lower bounds for sample complexity of finding an $\epsilo
n$-optimal policy of an infinite-horizon average-reward Markov decision process
(MDP) given access to a generative model. When the mixing time of the probabilit
y transition matrix of all policies is at most t_mix, we provide an a
lgorithm that solves the problem using $\widetilde{O}(t_\mathrm{mix} \epsilon^{-
3})$ (oblivious) samples per state-action pair. Further, we provide a lower boun
d showing that a linear dependence on t_mix is necessary in the worst
 case for any algorithm which computes oblivious samples. We obtain our results
by establishing connections between infinite-horizon average-reward MDPs and dis
counted MDPs of possible further utility.
****
Almost Optimal Anytime Algorithm for Batched Multi-Armed Bandits
Tianyuan Jin, Jing Tang, Pan Xu, Keke Huang, Xiaokui Xiao, Quanquan Gu
In batched multi-armed bandit problems, the learner can adaptively pull arms and
 adjust strategy in batches. In many real applications, not only the regret but
also the batch complexity need to be optimized. Existing batched bandit algorith
ms usually assume that the time horizon T is known in advance. However, many app
lications involve an unpredictable stopping time. In this paper, we study the an
ytime batched multi-armed bandit problem. We propose an anytime algorithm that a
chieves the asymptotically optimal regret for exponential families of reward dis
tributions with $O(\log \log T \ilog^{\alpha} (T))$ \footnote{Notation \ilog^{\a
lpha} (T) is the result of iteratively applying the logarithm function on T for
\alpha times, e.g., \ilog^{3} (T)=\log\log\log T.} batches, where $\alpha\in O_{
T}(1)$. Moreover, we prove that for any constant c>0, no algorithm can achieve t
he asymptotically optimal regret within c\log\log T batches.
****
MOTS: Minimax Optimal Thompson Sampling
Tianyuan Jin, Pan Xu, Jieming Shi, Xiaokui Xiao, Quanquan Gu
Thompson sampling is one of the most widely used algorithms in many online decis
ion problems due to its simplicity for implementation and superior empirical per
formance over other state-of-the-art methods. Despite its popularity and empiric
al success, it has remained an open problem whether Thompson sampling can achiev
e the minimax optimal regret O(\sqrt{TK}) for K-armed bandit problems, where T i
s the total time horizon. In this paper we fill this long open gap by proposing
a new Thompson sampling algorithm called MOTS that adaptively truncates the samp
ling result of the chosen arm at each time step. We prove that this simple varia
nt of Thompson sampling achieves the minimax optimal regret bound O(\sqrt{TK}) f
or finite time horizon T and also the asymptotic optimal regret bound when T g
rows to infinity as well. This is the first time that the minimax optimality of
multi-armed bandit problems has been attained by Thompson sampling type of algor
ithms.
****
Is Pessimism Provably Efficient for Offline RL?
Ying Jin, Zhuoran Yang, Zhaoran Wang
We study offline reinforcement learning (RL), which aims to learn an optimal pol
icy based on a dataset collected a priori. Due to the lack of further interactio
ns with the environment, offline RL suffers from the insufficient coverage of th
e dataset, which eludes most existing theoretical analysis. In this paper, we pr
opose a pessimistic variant of the value iteration algorithm (PEVI), which incor
porates an uncertainty quantifier as the penalty function. Such a penalty functi
on simply flips the sign of the bonus function for promoting exploration in onli
ne RL, which makes it easily implementable and compatible with general function
approximators. Without assuming the sufficient coverage of the dataset, we estab
lish a data-dependent upper bound on the suboptimality of PEVI for general Marko
v decision processes (MDPs). When specialized to linear MDPs, it matches the inf
ormation-theoretic lower bound up to multiplicative factors of the dimension and
 horizon. In other words, pessimism is not only provably efficient but also mini
max optimal. In particular, given the dataset, the learned policy serves as the
“best effort” among all policies, as no other policies can do better. Our theore

tical analysis identifies the critical role of pessimism in eliminating a notion of spurious correlation, which emerges from the “irrelevant” trajectories that are less covered by the dataset and not informative for the optimal policy.
****
Adversarial Option-Aware Hierarchical Imitation Learning
Mingxuan Jing, Wenbing Huang, Fuchun Sun, Xiaojian Ma, Tao Kong, Chuang Gan, Lei Li
It has been a challenge to learning skills for an agent from long-horizon unannotated demonstrations. Existing approaches like Hierarchical Imitation Learning(HIL) are prone to compounding errors or suboptimal solutions. In this paper, we propose Option-GAIL, a novel method to learn skills at long horizon. The key idea of Option-GAIL is modeling the task hierarchy by options and train the policy via generative adversarial optimization. In particular, we propose an Expectation-Maximization(EM)-style algorithm: an E-step that samples the options of expert conditioned on the current learned policy, and an M-step that updates the low- and high-level policies of agent simultaneously to minimize the newly proposed option-occupancy measurement between the expert and the agent. We theoretically prove the convergence of the proposed algorithm. Experiments show that Option-GAIL outperforms other counterparts consistently across a variety of tasks.
****
Discrete-Valued Latent Preference Matrix Estimation with Graph Side Information
Changhun Jo, Kangwook Lee
Incorporating graph side information into recommender systems has been widely used to better predict ratings, but relatively few works have focused on theoretical guarantees. Ahn et al. (2018) firstly characterized the optimal sample complexity in the presence of graph side information, but the results are limited due to strict, unrealistic assumptions made on the unknown latent preference matrix and the structure of user clusters. In this work, we propose a new model in which 1) the unknown latent preference matrix can have any discrete values, and 2) users can be clustered into multiple clusters, thereby relaxing the assumptions made in prior work. Under this new model, we fully characterize the optimal sample complexity and develop a computationally-efficient algorithm that matches the optimal sample complexity. Our algorithm is robust to model errors and outperforms the existing algorithms in terms of prediction performance on both synthetic and real data.
****
Provable Lipschitz Certification for Generative Models
Matt Jordan, Alex Dimakis
We present a scalable technique for upper bounding the Lipschitz constant of generative models. We relate this quantity to the maximal norm over the set of attainable vector-Jacobian products of a given generative model. We approximate this set by layerwise convex approximations using zonotopes. Our approach generalizes and improves upon prior work using zonotope transformers and we extend to Lipschitz estimation of neural networks with large output dimension. This provides efficient and tight bounds on small networks and can scale to generative models on VAE and DCGAN architectures.
****
Isometric Gaussian Process Latent Variable Model for Dissimilarity Data
Martin Jørgensen, Soren Hauberg
We present a probabilistic model where the latent variable respects both the distances and the topology of the modeled data. The model leverages the Riemannian geometry of the generated manifold to endow the latent space with a well-defined stochastic distance measure, which is modeled locally as Nakagami distributions. These stochastic distances are sought to be as similar as possible to observed distances along a neighborhood graph through a censoring process. The model is inferred by variational inference based on observations of pairwise distances. We demonstrate how the new model can encode invariances in the learned manifolds.
****
On the Generalization Power of Overfitted Two-Layer Neural Tangent Kernel Models
Peizhong Ju, Xiaojun Lin, Ness Shroff

In this paper, we study the generalization performance of min ℓ_2-norm over
fitting solutions for the neural tangent kernel (NTK) model of a two-layer neura
l network with ReLU activation that has no bias term. We show that, depending on
 the ground-truth function, the test error of overfitted NTK models exhibits cha
racteristics that are different from the "double-descent" of other overparameter
ized linear models with simple Fourier or Gaussian features. Specifically, for a
 class of learnable functions, we provide a new upper bound of the generalizatio
n error that approaches a small limiting value, even when the number of neurons
p approaches infinity. This limiting value further decreases with the number o
f training samples n. For functions outside of this class, we provide a lower
bound on the generalization error that does not diminish to zero even when n a
nd p are both large.
****
Improved Confidence Bounds for the Linear Logistic Model and Applications to Ban
dits
Kwang-Sung Jun, Lalit Jain, Blake Mason, Houssam Nassif
We propose improved fixed-design confidence bounds for the linear logistic model
. Our bounds significantly improve upon the state-of-the-art bound by Li et al.
(2017) via recent developments of the self-concordant analysis of the logistic l
oss (Faury et al., 2020). Specifically, our confidence bound avoids a direct dep
endence on $1/\kappa$, where κ is the minimal variance over all arms’ rew
ard distributions. In general, $1/\kappa$ scales exponentially with the norm of
the unknown linear parameter θ^*. Instead of relying on this worst case q
uantity, our confidence bound for the reward of any given arm depends directly o
n the variance of that arm’s reward distribution. We present two applications of
 our novel bounds to pure exploration and regret minimization logistic bandits i
mproving upon state-of-the-art performance guarantees. For pure exploration we a
lso provide a lower bound highlighting a dependence on $1/\kappa$ for a family o
f instances.
****
Detection of Signal in the Spiked Rectangular Models
Ji Hyung Jung, Hye Won Chung, Ji Oon Lee
We consider the problem of detecting signals in the rank-one signal-plus-noise d
ata matrix models that generalize the spiked Wishart matrices. We show that the
principal component analysis can be improved by pre-transforming the matrix entr
ies if the noise is non-Gaussian. As an intermediate step, we prove a sharp phas
e transition of the largest eigenvalues of spiked rectangular matrices, which ex
tends the Baik–Ben Arous–Péché (BBP) transition. We also propose a hypothesis te
st to detect the presence of signal with low computational complexity, based on
the linear spectral statistics, which minimizes the sum of the Type-I and Type-I
I errors when the noise is Gaussian.
****
Estimating Identifiable Causal Effects on Markov Equivalence Class through Doubl
e Machine Learning
Yonghan Jung, Jin Tian, Elias Bareinboim
General methods have been developed for estimating causal effects from observati
onal data under causal assumptions encoded in the form of a causal graph. Most o
f this literature assumes that the underlying causal graph is completely specifi
ed. However, only observational data is available in most practical settings, wh
ich means that one can learn at most a Markov equivalence class (MEC) of the und
erlying causal graph. In this paper, we study the problem of causal estimation f
rom a MEC represented by a partial ancestral graph (PAG), which is learnable fro
m observational data. We develop a general estimator for any identifiable causal
 effects in a PAG. The result fills a gap for an end-to-end solution to causal i
nference from observational data to effects estimation. Specifically, we develop
 a complete identification algorithm that derives an influence function for any
identifiable causal effects from PAGs. We then construct a double/debiased machi
ne learning (DML) estimator that is robust to model misspecification and biases
in nuisance function estimation, permitting the use of modern machine learning t
echniques. Simulation results corroborate with the theory.

****
A Nullspace Property for Subspace-Preserving Recovery
Mustafa D Kaba, Chong You, Daniel P Robinson, Enrique Mallada, Rene Vidal
Much of the theory for classical sparse recovery is based on conditions on the dictionary that are both necessary and sufficient (e.g., nullspace property) or only sufficient (e.g., incoherence and restricted isometry). In contrast, much of the theory for subspace-preserving recovery, the theoretical underpinnings for sparse subspace classification and clustering methods, is based on conditions on the subspaces and the data that are only sufficient (e.g., subspace incoherence and data inner-radius). This paper derives a necessary and sufficient condition for subspace-preserving recovery that is inspired by the classical nullspace property.Based on this novel condition, called here the subspace nullspace property, we derive equivalent characterizations that either admit a clear geometric interpretation that relates data distribution and subspace separation to the recovery success, or can be verified using a finite set of extreme points of a properly defined set. We further exploit these characterizations to derive new sufficient conditions, based on inner-radius and outer-radius measures and dual bounds, that generalize existing conditions and preserve the geometric interpretations. These results fill an important gap in the subspace-preserving recovery literature.
****
Training Recurrent Neural Networks via Forward Propagation Through Time
Anil Kag, Venkatesh Saligrama
Back-propagation through time (BPTT) has been widely used for training Recurrent Neural Networks (RNNs). BPTT updates RNN parameters on an instance by back-propagating the error in time over the entire sequence length, and as a result, leads to poor trainability due to the well-known gradient explosion/decay phenomena. While a number of prior works have proposed to mitigate vanishing/explosion effect through careful RNN architecture design, these RNN variants still train with BPTT. We propose a novel forward-propagation algorithm, FPTT, where at each time, for an instance, we update RNN parameters by optimizing an instantaneous risk function. Our proposed risk is a regularization penalty at time t that evolves dynamically based on previously observed losses, and allows for RNN parameter updates to converge to a stationary solution of the empirical RNN objective. We consider both sequence-to-sequence as well as terminal loss problems. Empirically FPTT outperforms BPTT on a number of well-known benchmark tasks, thus enabling architectures like LSTMs to solve long range dependencies problems.
****
The Distributed Discrete Gaussian Mechanism for Federated Learning with Secure Aggregation
Peter Kairouz, Ziyu Liu, Thomas Steinke
We consider training models on private data that are distributed across user devices. To ensure privacy, we add on-device noise and use secure aggregation so that only the noisy sum is revealed to the server. We present a comprehensive end-to-end system, which appropriately discretizes the data and adds discrete Gaussian noise before performing secure aggregation. We provide a novel privacy analysis for sums of discrete Gaussians and carefully analyze the effects of data quantization and modular summation arithmetic. Our theoretical guarantees highlight the complex tension between communication, privacy, and accuracy. Our extensive experimental results demonstrate that our solution is essentially able to match the accuracy to central differential privacy with less than 16 bits of precision per value.
****
Practical and Private (Deep) Learning Without Sampling or Shuffling
Peter Kairouz, Brendan Mcmahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, Zheng Xu
We consider training models with differential privacy (DP) using mini-batch gradients. The existing state-of-the-art, Differentially Private Stochastic Gradient Descent (DP-SGD), requires \emph{privacy amplification by sampling or shuffling} to obtain the best privacy/accuracy/computation trade-offs. Unfortunately, the

precise requirements on exact sampling and shuffling can be hard to obtain in important practical scenarios, particularly federated learning (FL). We design and analyze a DP variant of Follow-The-Regularized-Leader (DP-FTRL) that compares favorably (both theoretically and empirically) to amplified DP-SGD, while allowing for much more flexible data access patterns. DP-FTRL does not use any form of privacy amplification.
****
A Differentiable Point Process with Its Application to Spiking Neural Networks
Hiroshi Kajino
This paper is concerned about a learning algorithm for a probabilistic model of spiking neural networks (SNNs). Jimenez Rezende & Gerstner (2014) proposed a stochastic variational inference algorithm to train SNNs with hidden neurons. The algorithm updates the variational distribution using the score function gradient estimator, whose high variance often impedes the whole learning algorithm. This paper presents an alternative gradient estimator for SNNs based on the path-wise gradient estimator. The main technical difficulty is a lack of a general method to differentiate a realization of an arbitrary point process, which is necessary to derive the path-wise gradient estimator. We develop a differentiable point process, which is the technical highlight of this paper, and apply it to derive the path-wise gradient estimator for SNNs. We investigate the effectiveness of our gradient estimator through numerical simulation.
****
Projection techniques to update the truncated SVD of evolving matrices with applications
Vasileios Kalantzis, Georgios Kollias, Shashanka Ubaru, Athanasios N. Nikolakopoulos, Lior Horesh, Kenneth Clarkson
This submission considers the problem of updating the rank-k truncated Singular Value Decomposition (SVD) of matrices subject to the addition of new rows and/or columns over time. Such matrix problems represent an important computational kernel in applications such as Latent Semantic Indexing and Recommender Systems. Nonetheless, the proposed framework is purely algebraic and targets general updating problems. The algorithm presented in this paper undertakes a projection viewpoint and focuses on building a pair of subspaces which approximate the linear span of the sought singular vectors of the updated matrix. We discuss and analyze two different choices to form the projection subspaces. Results on matrices from real applications suggest that the proposed algorithm can lead to higher accuracy, especially for the singular triplets associated with the largest modulus singular values. Several practical details and key differences with other approaches are also discussed.
****
Optimal Off-Policy Evaluation from Multiple Logging Policies
Nathan Kallus, Yuta Saito, Masatoshi Uehara
We study off-policy evaluation (OPE) from multiple logging policies, each generating a dataset of fixed size, i.e., stratified sampling. Previous work noted that in this setting the ordering of the variances of different importance sampling estimators is instance-dependent, which brings up a dilemma as to which importance sampling weights to use. In this paper, we resolve this dilemma by finding the OPE estimator for multiple loggers with minimum variance for any instance, i.e., the efficient one. In particular, we establish the efficiency bound under stratified sampling and propose an estimator achieving this bound when given consistent q-estimates. To guard against misspecification of q-functions, we also provide a way to choose the control variate in a hypothesis class to minimize variance. Extensive experiments demonstrate the benefits of our methods' efficiently leveraging of the stratified sampling of off-policy data from multiple loggers.
****
Efficient Performance Bounds for Primal-Dual Reinforcement Learning from Demonstrations
Angeliki Kamoutsi, Goran Banjac, John Lygeros
We consider large-scale Markov decision processes with an unknown cost function

and address the problem of learning a policy from a finite set of expert demonst
rations. ■■■We assume that the learner is not allowed to interact with the exper
t and has no access to reinforcement signal of any kind. ■■■Existing inverse rei
nforcement learning methods come with strong theoretical guarantees, but are com
putationally expensive, while state-of-the-art policy optimization algorithms ac
hieve significant empirical success, but are hampered by limited theoretical und
erstanding. ■■■To bridge the gap between theory and practice, we introduce a nov
el bilinear saddle-point framework using Lagrangian duality. ■■■The proposed pri
mal-dual viewpoint allows us to develop a model-free provably efficient algorith
m through the lens of stochastic convex optimization. The method enjoys the adva
ntages of simplicity of implementation, low memory requirements, and computation
al and sample complexities independent of the number of states. We further prese
nt an equivalent no-regret online-learning interpretation.
****
Statistical Estimation from Dependent Data
Vardis Kandiros, Yuval Dagan, Nishanth Dikkala, Surbhi Goel, Constantinos Daskal
akis
We consider a general statistical estimation problem wherein binary labels acros
s different observations are not independent conditioning on their feature vecto
rs, but dependent, capturing settings where e.g. these observations are collecte
d on a spatial domain, a temporal domain, or a social network, which induce depe
ndencies. We model these dependencies in the language of Markov Random Fields an
d, importantly, allow these dependencies to be substantial, i.e. do not assume t
hat the Markov Random Field capturing these dependencies is in high temperature.
 As our main contribution we provide algorithms and statistically efficient esti
mation rates for this model, giving several instantiations of our bounds in logi
stic regression, sparse logistic regression, and neural network regression setti
ngs with dependent data. Our estimation guarantees follow from novel results for
 estimating the parameters (i.e. external fields and interaction strengths) of I
sing models from a single sample.
****
SKIing on Simplices: Kernel Interpolation on the Permutohedral Lattice for Scala
ble Gaussian Processes
Sanyam Kapoor, Marc Finzi, Ke Alexander Wang, Andrew Gordon Gordon Wilson
State-of-the-art methods for scalable Gaussian processes use iterative algorithm
s, requiring fast matrix vector multiplies (MVMs) with the co-variance kernel. T
he Structured Kernel Interpolation (SKI) framework accelerates these MVMs by per
forming efficient MVMs on a grid and interpolating back to the original space. I
n this work, we develop a connection between SKI and the permutohedral lattice u
sed for high-dimensional fast bilateral filtering. Using a sparse simplicial gri
d instead of a dense rectangular one, we can perform GP inference exponentially
faster in the dimension than SKI. Our approach, Simplex-GP, enables scaling SKI
to high dimensions, while maintaining strong predictive performance. We addition
ally provide a CUDA implementation of Simplex-GP, which enables significant GPU
acceleration of MVM based inference.
****
Variational Auto-Regressive Gaussian Processes for Continual Learning
Sanyam Kapoor, Theofanis Karaletsos, Thang D Bui
Through sequential construction of posteriors on observing data online, Bayes’ t
heorem provides a natural framework for continual learning. We develop Variation
al Auto-Regressive Gaussian Processes (VAR-GPs), a principled posterior updating
 mechanism to solve sequential tasks in continual learning. By relying on sparse
 inducing point approximations for scalable posteriors, we propose a novel auto-
regressive variational distribution which reveals two fruitful connections to ex
isting results in Bayesian inference, expectation propagation and orthogonal ind
ucing points. Mean predictive entropy estimates show VAR-GPs prevent catastrophi
c forgetting, which is empirically supported by strong performance on modern con
tinual learning benchmarks against competitive baselines. A thorough ablation st
udy demonstrates the efficacy of our modeling choices.
****

Off-Policy Confidence Sequences
Nikos Karampatziakis, Paul Mineiro, Aaditya Ramdas
We develop confidence bounds that hold uniformly over time for off-policy evaluation in the contextual bandit setting. These confidence sequences are based on recent ideas from martingale analysis and are non-asymptotic, non-parametric, and valid at arbitrary stopping times. We provide algorithms for computing these confidence sequences that strike a good balance between computational and statistical efficiency. We empirically demonstrate the tightness of our approach in terms of failure probability and width and apply it to the “gated deployment” problem of safely upgrading a production contextual bandit system.
****
Learning from History for Byzantine Robust Optimization
Sai Praneeth Karimireddy, Lie He, Martin Jaggi
Byzantine robustness has received significant attention recently given its importance for distributed and federated learning. In spite of this, we identify severe flaws in existing algorithms even when the data across the participants is identically distributed. First, we show realistic examples where current state of the art robust aggregation rules fail to converge even in the absence of any Byzantine attackers. Secondly, we prove that even if the aggregation rules may succeed in limiting the influence of the attackers in a single round, the attackers can couple their attacks across time eventually leading to divergence. To address these issues, we present two surprisingly simple strategies: a new robust iterative clipping procedure, and incorporating worker momentum to overcome time-coupled attacks. This is the first provably robust method for the standard stochastic optimization setting.
****
Non-Negative Bregman Divergence Minimization for Deep Direct Density Ratio Estimation
Masahiro Kato, Takeshi Teshima
Density ratio estimation (DRE) is at the core of various machine learning tasks such as anomaly detection and domain adaptation. In the DRE literature, existing studies have extensively studied methods based on Bregman divergence (BD) minimization. However, when we apply the BD minimization with highly flexible models, such as deep neural networks, it tends to suffer from what we call train-loss hacking, which is a source of over-fitting caused by a typical characteristic of empirical BD estimators. In this paper, to mitigate train-loss hacking, we propose non-negative correction for empirical BD estimators. Theoretically, we confirm the soundness of the proposed method through a generalization error bound. In our experiments, the proposed methods show favorable performances in inlier-based outlier detection.
****
Improved Algorithms for Agnostic Pool-based Active Classification
Julian Katz-Samuels, Jifan Zhang, Lalit Jain, Kevin Jamieson
We consider active learning for binary classification in the agnostic pool-based setting. The vast majority of works in active learning in the agnostic setting are inspired by the CAL algorithm where each query is uniformly sampled from the disagreement region of the current version space. The sample complexity of such algorithms is described by a quantity known as the disagreement coefficient which captures both the geometry of the hypothesis space as well as the underlying probability space. To date, the disagreement coefficient has been justified by minimax lower bounds only, leaving the door open for superior instance dependent sample complexities. In this work we propose an algorithm that, in contrast to uniform sampling over the disagreement region, solves an experimental design problem to determine a distribution over examples from which to request labels. We show that the new approach achieves sample complexity bounds that are never worse than the best disagreement coefficient-based bounds, but in specific cases can be dramatically smaller. From a practical perspective, the proposed algorithm requires no hyperparameters to tune (e.g., to control the aggressiveness of sampling), and is computationally efficient by means of assuming access to an empirical risk minimization oracle (without any constraints). Empirically, we demonstrat

e that our algorithm is superior to state of the art agnostic active learning al gorithms on image classification datasets.
****
When Does Data Augmentation Help With Membership Inference Attacks?
Yigitcan Kaya, Tudor Dumitras
Deep learning models often raise privacy concerns as they leak information about their training data. This leakage enables membership inference attacks (MIA) th at can identify whether a data point was in a model’s training set. Research sho ws that some ’data augmentation’ mechanisms may reduce the risk by combatting a key factor increasing the leakage, overfitting. While many mechanisms exist, the ir effectiveness against MIAs and privacy properties have not been studied syste matically. Employing two recent MIAs, we explore the lower bound on the risk in the absence of formal upper bounds. First, we evaluate 7 mechanisms and differen tial privacy, on three image classification tasks. We find that applying augment ation to increase the model’s utility does not mitigate the risk and protection comes with a utility penalty. Further, we also investigate why popular label smo othing mechanism consistently amplifies the risk. Finally, we propose ’loss-rank -correlation’ (LRC) metric to assess how similar the effects of different mechan isms are. This, for example, reveals the similarity of applying high-intensity a ugmentation against MIAs to simply reducing the training time. Our findings emph asize the utility-privacy trade-off and provide practical guidelines on using au gmentation to manage the trade-off.
****
Regularized Submodular Maximization at Scale
Ehsan Kazemi, Shervin Minaee, Moran Feldman, Amin Karbasi
In this paper, we propose scalable methods for maximizing a regularized submodul ar function $f \triangleq g-\ell$ expressed as the difference between a monotone submodular function g and a modular function ℓ. Submodularity is inheren tly related to the notions of diversity, coverage, and representativeness. In pa rticular, finding the mode (i.e., the most likely configuration) of many popular probabilistic models of diversity, such as determinantal point processes and st rongly log-concave distributions, involves maximization of (regularized) submodu lar functions. Since a regularized function f can potentially take on negative values, the classic theory of submodular maximization, which heavily relies on the non-negativity assumption of submodular functions, is not applicable. To cir cumvent this challenge, we develop the first one-pass streaming algorithm for ma ximizing a regularized submodular function subject to a k-cardinality constrai nt. Furthermore, we develop the first distributed algorithm that returns a solut ion S in $O(1/ \epsilon)$ rounds of MapReduce computation. We highlight that o ur result, even for the unregularized case where the modular term ℓ is zero , improves the memory and communication complexity of the state-of-the-art by a factor of $O(1/ \epsilon)$ while arguably provides a simpler distributed algorit hm and a unifying analysis. We empirically study the performance of our scalable methods on a set of real-life applications, including finding the mode of negat ively correlated distributions, vertex cover of social networks, and several dat a summarization tasks.
****
Prior Image-Constrained Reconstruction using Style-Based Generative Models
Varun A Kelkar, Mark Anastasio
Obtaining a useful estimate of an object from highly incomplete imaging measurem ents remains a holy grail of imaging science. Deep learning methods have shown p romise in learning object priors or constraints to improve the conditioning of a n ill-posed imaging inverse problem. In this study, a framework for estimating a n object of interest that is semantically related to a known prior image, is pro posed. An optimization problem is formulated in the disentangled latent space of a style-based generative model, and semantically meaningful constraints are imp osed using the disentangled latent representation of the prior image. Stable rec overy from incomplete measurements with the help of a prior image is theoretical ly analyzed. Numerical experiments demonstrating the superior performance of our approach as compared to related methods are presented.

****
Self Normalizing Flows
Thomas A Keller, Jorn W.T. Peters, Priyank Jaini, Emiel Hoogeboom, Patrick Forré
, Max Welling
Efficient gradient computation of the Jacobian determinant term is a core proble
m in many machine learning settings, and especially so in the normalizing flow f
ramework. Most proposed flow models therefore either restrict to a function clas
s with easy evaluation of the Jacobian determinant, or an efficient estimator th
ereof. However, these restrictions limit the performance of such density models,
 frequently requiring significant depth to reach desired performance levels. In
this work, we propose \emph{Self Normalizing Flows}, a flexible framework for tr
aining normalizing flows by replacing expensive terms in the gradient by learned
 approximate inverses at each layer. This reduces the computational complexity o
f each layer’s exact update from $\mathcal{O}(D^3)$ to $\mathcal{O}(D^2)$, allow
ing for the training of flow architectures which were otherwise computationally
infeasible, while also providing efficient sampling. We show experimentally that
 such models are remarkably stable and optimize to similar data likelihood value
s as their exact gradient counterparts, while training more quickly and surpassi
ng the performance of functionally constrained counterparts.
****
Interpretable Stability Bounds for Spectral Graph Filters
Henry Kenlay, Dorina Thanou, Xiaowen Dong
Graph-structured data arise in a variety of real-world context ranging from sens
or and transportation to biological and social networks. As a ubiquitous tool to
 process graph-structured data, spectral graph filters have been used to solve c
ommon tasks such as denoising and anomaly detection, as well as design deep lear
ning architectures such as graph neural networks. Despite being an important too
l, there is a lack of theoretical understanding of the stability properties of s
pectral graph filters, which are important for designing robust machine learning
 models. In this paper, we study filter stability and provide a novel and interp
retable upper bound on the change of filter output, where the bound is expressed
 in terms of the endpoint degrees of the deleted and newly added edges, as well
as the spatial proximity of those edges. This upper bound allows us to reason, i
n terms of structural properties of the graph, when a spectral graph filter will
 be stable. We further perform extensive experiments to verify intuition that ca
n be gained from the bound.
****
Affine Invariant Analysis of Frank-Wolfe on Strongly Convex Sets
Thomas Kerdreux, Lewis Liu, Simon Lacoste-Julien, Damien Scieur
It is known that the Frank-Wolfe (FW) algorithm, which is affine covariant, enjo
ys faster convergence rates than $\mathcal{O}\left(1/K\right)$ when the constrai
nt set is strongly convex. However, these results rely on norm-dependent assumpt
ions, usually incurring non-affine invariant bounds, in contradiction with FW’s
affine covariant property. In this work, we introduce new structural assumptions
 on the problem (such as the directional smoothness) and derive an affine invari
ant, norm-independent analysis of Frank-Wolfe. We show that our rates are better
 than any other known convergence rates of FW in this setting. Based on our anal
ysis, we propose an affine invariant backtracking line-search. Interestingly, we
 show that typical backtracking line-searches using smoothness of the objective
function present similar performances than its affine invariant counterpart, des
pite using affine dependent norms in the step size’s computation.
****
Markpainting: Adversarial Machine Learning meets Inpainting
David Khachaturov, Ilia Shumailov, Yiren Zhao, Nicolas Papernot, Ross Anderson
Inpainting is a learned interpolation technique that is based on generative mode
ling and used to populate masked or missing pieces in an image; it has wide appl
ications in picture editing and retouching. Recently, inpainting started being u
sed for watermark removal, raising concerns. In this paper we study how to manip
ulate it using our markpainting technique. First, we show how an image owner wit
h access to an inpainting model can augment their image in such a way that any a

ttempt to edit it using that model will add arbitrary visible information. We find that we can target multiple different models simultaneously with our technique. This can be designed to reconstitute a watermark if the editor had been trying to remove it. Second, we show that our markpainting technique is transferable to models that have different architectures or were trained on different datasets, so watermarks created using it are difficult for adversaries to remove. Markpainting is novel and can be used as a manipulation alarm that becomes visible in the event of inpainting. Source code is available at: https://github.com/iliaishacked/markpainting.

****

Finite-Sample Analysis of Off-Policy Natural Actor-Critic Algorithm

Sajad Khodadadian, Zaiwei Chen, Siva Theja Maguluri

In this paper, we provide finite-sample convergence guarantees for an off-policy variant of the natural actor-critic (NAC) algorithm based on Importance Sampling. In particular, we show that the algorithm converges to a global optimal policy with a sample complexity of $\mathcal{O}(\epsilon^{-3}\log^2(1/\epsilon))$ under an appropriate choice of stepsizes. In order to overcome the issue of large variance due to Importance Sampling, we propose the Q-trace algorithm for the critic, which is inspired by the V-trace algorithm (Espeholt et al., 2018). This enables us to explicitly control the bias and variance, and characterize the trade-off between them. As an advantage of off-policy sampling, a major feature of our result is that we do not need any additional assumptions, beyond the ergodicity of the Markov chain induced by the behavior policy.

****

Functional Space Analysis of Local GAN Convergence

Valentin Khrulkov, Artem Babenko, Ivan Oseledets

Recent work demonstrated the benefits of studying continuous-time dynamics governing the GAN training. However, this dynamics is analyzed in the model parameter space, which results in finite-dimensional dynamical systems. We propose a novel perspective where we study the local dynamics of adversarial training in the general functional space and show how it can be represented as a system of partial differential equations. Thus, the convergence properties can be inferred from the eigenvalues of the resulting differential operator. We show that these eigenvalues can be efficiently estimated from the target dataset before training. Our perspective reveals several insights on the practical tricks commonly used to stabilize GANs, such as gradient penalty, data augmentation, and advanced integration schemes. As an immediate practical benefit, we demonstrate how one can a priori select an optimal data augmentation strategy for a particular generation task.

****

"Hey, that’s not an ODE": Faster ODE Adjoints via Seminorms

Patrick Kidger, Ricky T. Q. Chen, Terry J Lyons

Neural differential equations may be trained by backpropagating gradients via the adjoint method, which is another differential equation typically solved using an adaptive-step-size numerical differential equation solver. A proposed step is accepted if its error, \emph{relative to some norm}, is sufficiently small; else it is rejected, the step is shrunk, and the process is repeated. Here, we demonstrate that the particular structure of the adjoint equations makes the usual choices of norm (such as L^2) unnecessarily stringent. By replacing it with a more appropriate (semi)norm, fewer steps are unnecessarily rejected and the backpropagation is made faster. This requires only minor code modifications. Experiments on a wide range of tasks—including time series, generative modeling, and physical control—demonstrate a median improvement of 40% fewer function evaluations. On some problems we see as much as 62% fewer function evaluations, so that the overall training time is roughly halved.

****

Neural SDEs as Infinite-Dimensional GANs

Patrick Kidger, James Foster, Xuechen Li, Terry J Lyons

Stochastic differential equations (SDEs) are a staple of mathematical modelling of temporal dynamics. However, a fundamental limitation has been that such model

s have typically been relatively inflexible, which recent work introducing Neura
l SDEs has sought to solve. Here, we show that the current classical approach to
 fitting SDEs may be approached as a special case of (Wasserstein) GANs, and in
doing so the neural and classical regimes may be brought together. The input noi
se is Brownian motion, the output samples are time-evolving paths produced by a
numerical solver, and by parameterising a discriminator as a Neural Controlled D
ifferential Equation (CDE), we obtain Neural SDEs as (in modern machine learning
 parlance) continuous-time generative time series models. Unlike previous work o
n this problem, this is a direct extension of the classical approach without ref
erence to either prespecified statistics or density functions. Arbitrary drift a
nd diffusions are admissible, so as the Wasserstein loss has a unique global min
ima, in the infinite data limit \textit{any} SDE may be learnt.
****
GRAD-MATCH: Gradient Matching based Data Subset Selection for Efficient Deep Mod
el Training
Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, Rishabh Iyer
The great success of modern machine learning models on large datasets is conting
ent on extensive computational resources with high financial and environmental c
osts. One way to address this is by extracting subsets that generalize on par wi
th the full data. In this work, we propose a general framework, GRAD-MATCH, whic
h finds subsets that closely match the gradient of the \emph{training or validat
ion} set. We find such subsets effectively using an orthogonal matching pursuit
algorithm. We show rigorous theoretical and convergence guarantees of the propos
ed algorithm and, through our extensive experiments on real-world datasets, show
 the effectiveness of our proposed framework. We show that GRAD-MATCH significan
tly and consistently outperforms several recent data-selection algorithms and ac
hieves the best accuracy-efficiency trade-off. GRAD-MATCH is available as a part
 of the CORDS toolkit: \url{https://github.com/decile-team/cords}.
****
Improving Predictors via Combination Across Diverse Task Categories
Kwang In Kim
Predictor combination is the problem of improving a task predictor using predict
ors of other tasks when the forms of individual predictors are unknown. Previous
 work approached this problem by nonparametrically assessing predictor relations
hips based on their joint evaluations on a shared sample. This limits their appl
ication to cases where all predictors are defined on the same task category, e.g
. all predictors estimate attributes of shoes. We present a new predictor combin
ation algorithm that overcomes this limitation. Our algorithm aligns the heterog
eneous domains of different predictors in a shared latent space to facilitate co
mparisons of predictors independently of the domains on which they are originall
y defined. We facilitate this by a new data alignment scheme that matches data d
istributions across task categories. Based on visual attribute ranking experimen
ts on datasets that span diverse task categories (e.g. shoes and animals), we de
monstrate that our approach often significantly improves the performances of the
 initial predictors.
****
Self-Improved Retrosynthetic Planning
Junsu Kim, Sungsoo Ahn, Hankook Lee, Jinwoo Shin
Retrosynthetic planning is a fundamental problem in chemistry for finding a path
way of reactions to synthesize a target molecule. Recently, search algorithms ha
ve shown promising results for solving this problem by using deep neural network
s (DNNs) to expand their candidate solutions, i.e., adding new reactions to reac
tion pathways. However, the existing works on this line are suboptimal; the retr
osynthetic planning problem requires the reaction pathways to be (a) represented
 by real-world reactions and (b) executable using “building block” molecules, ye
t the DNNs expand reaction pathways without fully incorporating such requirement
s. Motivated by this, we propose an end-to-end framework for directly training t
he DNNs towards generating reaction pathways with the desirable properties. Our
main idea is based on a self-improving procedure that trains the model to imitat
e successful trajectories found by itself. We also propose a novel reaction augm

entation scheme based on a forward reaction model. Our experiments demonstrate that our scheme significantly improves the success rate of solving the retrosynthetic problem from 86.84% to 96.32% while maintaining the performance of DNN for predicting valid reactions.
****
Reward Identification in Inverse Reinforcement Learning
Kuno Kim, Shivam Garg, Kirankumar Shiragur, Stefano Ermon
We study the problem of reward identifiability in the context of Inverse Reinforcement Learning (IRL). The reward identifiability question is critical to answer when reasoning about the effectiveness of using Markov Decision Processes (MDPs) as computational models of real world decision makers in order to understand complex decision making behavior and perform counterfactual reasoning. While identifiability has been acknowledged as a fundamental theoretical question in IRL, little is known about the types of MDPs for which rewards are identifiable, or even if there exist such MDPs. In this work, we formalize the reward identification problem in IRL and study how identifiability relates to properties of the MDP model. For deterministic MDP models with the MaxEntRL objective, we prove necessary and sufficient conditions for identifiability. Building on these results, we present efficient algorithms for testing whether or not an MDP model is identifiable.
****
I-BERT: Integer-only BERT Quantization
Sehoon Kim, Amir Gholami, Zhewei Yao, Michael W. Mahoney, Kurt Keutzer
Transformer based models, like BERT and RoBERTa, have achieved state-of-the-art results in many Natural Language Processing tasks. However, their memory footprint, inference latency, and power consumption are prohibitive efficient inference at the edge, and even at the data center. While quantization can be a viable solution for this, previous work on quantizing Transformer based models use floating-point arithmetic during inference, which cannot efficiently utilize integer-only logical units such as the recent Turing Tensor Cores, or traditional integer-only ARM processors. In this work, we propose I-BERT, a novel quantization scheme for Transformer based models that quantizes the entire inference with integer-only arithmetic. Based on lightweight integer-only approximation methods for nonlinear operations, e.g., GELU, Softmax, and Layer Normalization, I-BERT performs an end-to-end integer-only BERT inference without any floating point calculation. We evaluate our approach on GLUE downstream tasks using RoBERTa-Base/Large. We show that for both cases, I-BERT achieves similar (and slightly higher) accuracy as compared to the full-precision baseline. Furthermore, our preliminary implementation of I-BERT shows a speedup of 2.4- 4.0x for INT8 inference on a T4 GPU system as compared to FP32 inference. The framework has been developed in PyTorch and has been open-sourced.
****
Message Passing Adaptive Resonance Theory for Online Active Semi-supervised Learning
Taehyeong Kim, Injune Hwang, Hyundo Lee, Hyunseo Kim, Won-Seok Choi, Joseph J Lim, Byoung-Tak Zhang
Active learning is widely used to reduce labeling effort and training time by repeatedly querying only the most beneficial samples from unlabeled data. In real-world problems where data cannot be stored indefinitely due to limited storage or privacy issues, the query selection and the model update should be performed as soon as a new data sample is observed. Various online active learning methods have been studied to deal with these challenges; however, there are difficulties in selecting representative query samples and updating the model efficiently without forgetting. In this study, we propose Message Passing Adaptive Resonance Theory (MPART) that learns the distribution and topology of input data online. Through message passing on the topological graph, MPART actively queries informative and representative samples, and continuously improves the classification performance using both labeled and unlabeled data. We evaluate our model in stream-based selective sampling scenarios with comparable query selection strategies, showing that MPART significantly outperforms competitive models.

****
Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech
Jaehyeon Kim, Jungil Kong, Juhee Son
Several recent end-to-end text-to-speech (TTS) models enabling single-stage training and parallel sampling have been proposed, but their sample quality does not match that of two-stage TTS systems. In this work, we present a parallel end-to-end TTS method that generates more natural sounding audio than current two-stage models. Our method adopts variational inference augmented with normalizing flows and an adversarial training process, which improves the expressive power of generative modeling. We also propose a stochastic duration predictor to synthesize speech with diverse rhythms from input text. With the uncertainty modeling over latent variables and the stochastic duration predictor, our method expresses the natural one-to-many relationship in which a text input can be spoken in multiple ways with different pitches and rhythms. A subjective human evaluation (mean opinion score, or MOS) on the LJ Speech, a single speaker dataset, shows that our method outperforms the best publicly available TTS systems and achieves a MOS comparable to ground truth.
****
A Policy Gradient Algorithm for Learning to Learn in Multiagent Reinforcement Learning
Dong Ki Kim, Miao Liu, Matthew D Riemer, Chuangchuang Sun, Marwa Abdulhai, Golnaz Habibi, Sebastian Lopez-Cot, Gerald Tesauro, Jonathan How
A fundamental challenge in multiagent reinforcement learning is to learn beneficial behaviors in a shared environment with other simultaneously learning agents. In particular, each agent perceives the environment as effectively non-stationary due to the changing policies of other agents. Moreover, each agent is itself constantly learning, leading to natural non-stationarity in the distribution of experiences encountered. In this paper, we propose a novel meta-multiagent policy gradient theorem that directly accounts for the non-stationary policy dynamics inherent to multiagent learning settings. This is achieved by modeling our gradient updates to consider both an agent’s own non-stationary policy dynamics and the non-stationary policy dynamics of other agents in the environment. We show that our theoretically grounded approach provides a general solution to the multiagent learning problem, which inherently comprises all key aspects of previous state of the art approaches on this topic. We test our method on a diverse suite of multiagent benchmarks and demonstrate a more efficient ability to adapt to new agents as they learn than baseline methods across the full spectrum of mixed incentive, competitive, and cooperative domains.
****
Inferring Latent Dynamics Underlying Neural Population Activity via Neural Differential Equations
Timothy D. Kim, Thomas Z. Luo, Jonathan W. Pillow, Carlos D. Brody
An important problem in systems neuroscience is to identify the latent dynamics underlying neural population activity. Here we address this problem by introducing a low-dimensional nonlinear model for latent neural population dynamics using neural ordinary differential equations (neural ODEs), with noisy sensory inputs and Poisson spike train outputs. We refer to this as the Poisson Latent Neural Differential Equations (PLNDE) model. We apply the PLNDE framework to a variety of synthetic datasets, and show that it accurately infers the phase portraits and fixed points of nonlinear systems augmented to produce spike train data, including the FitzHugh-Nagumo oscillator, a 3-dimensional nonlinear spiral, and a nonlinear sensory decision-making model with attractor dynamics. Our model significantly outperforms existing methods at inferring single-trial neural firing rates and the corresponding latent trajectories that generated them, especially in the regime where the spike counts and number of trials are low. We then apply our model to multi-region neural population recordings from medial frontal cortex of rats performing an auditory decision-making task. Our model provides a general, interpretable framework for investigating the neural mechanisms of decision-making and other cognitive computations through the lens of dynamical systems.

****
The Lipschitz Constant of Self-Attention
Hyunjik Kim, George Papamakarios, Andriy Mnih
Lipschitz constants of neural networks have been explored in various contexts in deep learning, such as provable adversarial robustness, estimating Wasserstein distance, stabilising training of GANs, and formulating invertible neural networks. Such works have focused on bounding the Lipschitz constant of fully connected or convolutional networks, composed of linear maps and pointwise non-linearities. In this paper, we investigate the Lipschitz constant of self-attention, a non-linear neural network module widely used in sequence modelling. We prove that the standard dot-product self-attention is not Lipschitz for unbounded input domain, and propose an alternative L2 self-attention that is Lipschitz. We derive an upper bound on the Lipschitz constant of L2 self-attention and provide empirical evidence for its asymptotic tightness. To demonstrate the practical relevance of our theoretical work, we formulate invertible self-attention and use it in a Transformer-based architecture for a character-level language modelling task.
****
Unsupervised Skill Discovery with Bottleneck Option Learning
Jaekyeom Kim, Seohong Park, Gunhee Kim
Having the ability to acquire inherent skills from environments without any external rewards or supervision like humans is an important problem. We propose a novel unsupervised skill discovery method named Information Bottleneck Option Learning (IBOL). On top of the linearization of environments that promotes more various and distant state transitions, IBOL enables the discovery of diverse skills. It provides the abstraction of the skills learned with the information bottleneck framework for the options with improved stability and encouraged disentanglement. We empirically demonstrate that IBOL outperforms multiple state-of-the-art unsupervised skill discovery methods on the information-theoretic evaluations and downstream tasks in MuJoCo environments, including Ant, HalfCheetah, Hopper and D’Kitty. Our code is available at https://vision.snu.ac.kr/projects/ibol.
****
ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision
Wonjae Kim, Bokyung Son, Ildoo Kim
Vision-and-Language Pre-training (VLP) has improved performance on various joint vision-and-language downstream tasks. Current approaches to VLP heavily rely on image feature extraction processes, most of which involve region supervision (e.g., object detection) and the convolutional architecture (e.g., ResNet). Although disregarded in the literature, we find it problematic in terms of both (1) efficiency/speed, that simply extracting input features requires much more computation than the multimodal interaction steps; and (2) expressive power, as it is upper bounded to the expressive power of the visual embedder and its predefined visual vocabulary. In this paper, we present a minimal VLP model, Vision-and-Language Transformer (ViLT), monolithic in the sense that the processing of visual inputs is drastically simplified to just the same convolution-free manner that we process textual inputs. We show that ViLT is up to tens of times faster than previous VLP models, yet with competitive or better downstream task performance. Our code and pre-trained weights are available at https://github.com/dandelin/vilt.
****
Bias-Robust Bayesian Optimization via Dueling Bandits
Johannes Kirschner, Andreas Krause
We consider Bayesian optimization in settings where observations can be adversarially biased, for example by an uncontrolled hidden confounder. Our first contribution is a reduction of the confounded setting to the dueling bandit model. Then we propose a novel approach for dueling bandits based on information-directed sampling (IDS). Thereby, we obtain the first efficient kernelized algorithm for dueling bandits that comes with cumulative regret guarantees. Our analysis further generalizes a previously proposed semi-parametric linear bandit model to non-linear reward functions, and uncovers interesting links to doubly-robust estimation.

****
CLOCS: Contrastive Learning of Cardiac Signals Across Space, Time, and Patients
Dani Kiyasseh, Tingting Zhu, David A Clifton
The healthcare industry generates troves of unlabelled physiological data. This data can be exploited via contrastive learning, a self-supervised pre-training method that encourages representations of instances to be similar to one another. We propose a family of contrastive learning methods, CLOCS, that encourages representations across space, time, \textit{and} patients to be similar to one another. We show that CLOCS consistently outperforms the state-of-the-art methods, BYOL and SimCLR, when performing a linear evaluation of, and fine-tuning on, downstream tasks. We also show that CLOCS achieves strong generalization performance with only 25% of labelled training data. Furthermore, our training procedure naturally generates patient-specific representations that can be used to quantify patient-similarity.
****
Scalable Optimal Transport in High Dimensions for Graph Distances, Embedding Alignment, and More
Johannes Gasteiger, Marten Lienen, Stephan Günnemann
The current best practice for computing optimal transport (OT) is via entropy regularization and Sinkhorn iterations. This algorithm runs in quadratic time as it requires the full pairwise cost matrix, which is prohibitively expensive for large sets of objects. In this work we propose two effective log-linear time approximations of the cost matrix: First, a sparse approximation based on locality sensitive hashing (LSH) and, second, a Nystr{ö}m approximation with LSH-based sparse corrections, which we call locally corrected Nystr{ö}m (LCN). These approximations enable general log-linear time algorithms for entropy-regularized OT that perform well even for the complex, high-dimensional spaces common in deep learning. We analyse these approximations theoretically and evaluate them experimentally both directly and end-to-end as a component for real-world applications. Using our approximations for unsupervised word embedding alignment enables us to speed up a state-of-the-art method by a factor of 3 while also improving the accuracy by 3.1 percentage points without any additional model changes. For graph distance regression we propose the graph transport network (GTN), which combines graph neural networks (GNNs) with enhanced Sinkhorn. GTN outcompetes previous models by 48% and still scales log-linearly in the number of nodes.
****
Representational aspects of depth and conditioning in normalizing flows
Frederic Koehler, Viraj Mehta, Andrej Risteski
Normalizing flows are among the most popular paradigms in generative modeling, especially for images, primarily because we can efficiently evaluate the likelihood of a data point. This is desirable both for evaluating the fit of a model, and for ease of training, as maximizing the likelihood can be done by gradient descent. However, training normalizing flows comes with difficulties as well: models which produce good samples typically need to be extremely deep – which comes with accompanying vanishing/exploding gradient problems. A very related problem is that they are often poorly \emph{conditioned}: since they are parametrized as invertible maps from $\mathbb{R}^d \to \mathbb{R}^d$, and typical training data like images intuitively is lower-dimensional, the learned maps often have Jacobians that are close to being singular. In our paper, we tackle representational aspects around depth and conditioning of normalizing flows: both for general invertible architectures, and for a particular common architecture, affine couplings. We prove that $\Theta(1)$ affine coupling layers suffice to exactly represent a permutation or 1×1 convolution, as used in GLOW, showing that representationally the choice of partition is not a bottleneck for depth. We also show that shallow affine coupling networks are universal approximators in Wasserstein distance if ill-conditioning is allowed, and experimentally investigate related phenomena involving padding. Finally, we show a depth lower bound for general flow architectures with few neurons per layer and bounded Lipschitz constant.
****
WILDS: A Benchmark of in-the-Wild Distribution Shifts

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, Percy Liang
Distribution shifts—where the training distribution differs from the test distribution—can substantially degrade the accuracy of machine learning (ML) systems deployed in the wild. Despite their ubiquity in the real-world deployments, these distribution shifts are under-represented in the datasets widely used in the ML community today. To address this gap, we present WILDS, a curated benchmark of 10 datasets reflecting a diverse range of distribution shifts that naturally arise in real-world applications, such as shifts across hospitals for tumor identification; across camera traps for wildlife monitoring; and across time and location in satellite imaging and poverty mapping. On each dataset, we show that standard training yields substantially lower out-of-distribution than in-distribution performance. This gap remains even with models trained by existing methods for tackling distribution shifts, underscoring the need for new methods for training models that are more robust to the types of distribution shifts that arise in practice. To facilitate method development, we provide an open-source package that automates dataset loading, contains default model architectures and hyperparameters, and standardizes evaluations. The full paper, code, and leaderboards are available at https://wilds.stanford.edu.
****
One-sided Frank-Wolfe algorithms for saddle problems
Vladimir Kolmogorov, Thomas Pock
We study a class of convex-concave saddle-point problems of the form $\min_x\max_y ■Kx,y■+f_{\cal P}(x)-h^*(y)$ where K is a linear operator, $f_{\cal P}$ is the sum of a convex function f with a Lipschitz-continuous gradient and the indicator function of a bounded convex polytope ${\cal P}$, and $h^\ast$ is a convex (possibly nonsmooth) function. Such problem arises, for example, as a Lagrangian relaxation of various discrete optimization problems. Our main assumptions are the existence of an efficient {\em linear minimization oracle} (lmo) for $f_{\cal P}$ and an efficient {\em proximal map} ($prox$) for h^* which motivate the solution via a blend of proximal primal-dual algorithms and Frank-Wolfe algorithms. In case h^* is the indicator function of a linear constraint and function f is quadratic, we show a $O(1/n^2)$ convergence rate on the dual objective, requiring $O(n \log n)$ calls of lmo. If the problem comes from the constrained optimization problem $\min_{x\in\mathbb R^d}\{f_{\cal P}(x)\:|\:Ax-b=0\}$ then we additionally get bound $O(1/n^2)$ both on the primal gap and on the infeasibility gap. In the most general case, we show a $O(1/n)$ convergence rate of the primal-dual gap again requiring $O(n\log n)$ calls of lmo. To the best of our knowledge, this improves on the known convergence rates for the considered class of saddle-point problems. We show applications to labeling problems frequently appearing in machine learning and computer vision.
****
A Lower Bound for the Sample Complexity of Inverse Reinforcement Learning
Abi Komanduru, Jean Honorio
Inverse reinforcement learning (IRL) is the task of finding a reward function that generates a desired optimal policy for a given Markov Decision Process (MDP). This paper develops an information-theoretic lower bound for the sample complexity of the finite state, finite action IRL problem. A geometric construction of β-strict separable IRL problems using spherical codes is considered. Properties of the ensemble size as well as the Kullback-Leibler divergence between the generated trajectories are derived. The resulting ensemble is then used along with Fano’s inequality to derive a sample complexity lower bound of $O(n \log n)$, where n is the number of states in the MDP.
****
Consensus Control for Decentralized Deep Learning
Lingjing Kong, Tao Lin, Anastasia Koloskova, Martin Jaggi, Sebastian Stich
Decentralized training of deep learning models enables on-device learning over n

etworks, as well as efficient scaling to large compute clusters. Experiments in earlier works reveal that, even in a data-center setup, decentralized training often suffers from the degradation in the quality of the model: the training and test performance of models trained in a decentralized fashion is in general worse than that of models trained in a centralized fashion, and this performance drop is impacted by parameters such as network size, communication topology and data partitioning. We identify the changing consensus distance between devices as a key parameter to explain the gap between centralized and decentralized training. We show in theory that when the training consensus distance is lower than a critical quantity, decentralized training converges as fast as the centralized counterpart. We empirically validate that the relation between generalization performance and consensus distance is consistent with this theoretical observation. Our empirical insights allow the principled design of better decentralized training schemes that mitigate the performance drop. To this end, we provide practical training guidelines and exemplify its effectiveness on the data-center setup as the important first step.
****
A Distribution-dependent Analysis of Meta Learning
Mikhail Konobeev, Ilja Kuzborskij, Csaba Szepesvari
A key problem in the theory of meta-learning is to understand how the task distributions influence transfer risk, the expected error of a meta-learner on a new task drawn from the unknown task distribution. In this paper, focusing on fixed design linear regression with Gaussian noise and a Gaussian task (or parameter) distribution, we give distribution-dependent lower bounds on the transfer risk of any algorithm, while we also show that a novel, weighted version of the so-called biased regularized regression method is able to match these lower bounds up to a fixed constant factor. Notably, the weighting is derived from the covariance of the Gaussian task distribution. Altogether, our results provide a precise characterization of the difficulty of meta-learning in this Gaussian setting. While this problem setting may appear simple, we show that it is rich enough to unify the “parameter sharing” and “representation learning” streams of meta-learning; in particular, representation learning is obtained as the special case when the covariance matrix of the task distribution is unknown. For this case we propose to adopt the EM method, which is shown to enjoy efficient updates in our case. The paper is completed by an empirical study of EM. In particular, our experimental results show that the EM algorithm can attain the lower bound as the number of tasks grows, while the algorithm is also successful in competing with its alternatives when used in a representation learning context.
****
Evaluating Robustness of Predictive Uncertainty Estimation: Are Dirichlet-based Models Reliable?
Anna-Kathrin Kopetzki, Bertrand Charpentier, Daniel Zügner, Sandhya Giri, Stephan Günnemann
Dirichlet-based uncertainty (DBU) models are a recent and promising class of uncertainty-aware models. DBU models predict the parameters of a Dirichlet distribution to provide fast, high-quality uncertainty estimates alongside with class predictions. In this work, we present the first large-scale, in-depth study of the robustness of DBU models under adversarial attacks. Our results suggest that uncertainty estimates of DBU models are not robust w.r.t. three important tasks: (1) indicating correctly and wrongly classified samples; (2) detecting adversarial examples; and (3) distinguishing between in-distribution (ID) and out-of-distribution (OOD) data. Additionally, we explore the first approaches to make DBU mod- els more robust. While adversarial training has a minor effect, our median smoothing based ap- proach significantly increases robustness of DBU models.
****
Kernel Stein Discrepancy Descent
Anna Korba, Pierre-Cyril Aubin-Frankowski, Szymon Majewski, Pierre Ablin
Among dissimilarities between probability distributions, the Kernel Stein Discrepancy (KSD) has received much interest recently. We investigate the properties of its Wasserstein gradient flow to approximate a target probability distribution

π on $\mathbb{R}^d$, known up to a normalization constant. This leads to a straightforwardly implementable, deterministic score-based method to sample from π, named KSD Descent, which uses a set of particles to approximate π. Remarkably, owing to a tractable loss function, KSD Descent can leverage robust parameter-free optimization schemes such as L-BFGS; this contrasts with other popular particle-based schemes such as the Stein Variational Gradient Descent algorithm. We study the convergence properties of KSD Descent and demonstrate its practical relevance. However, we also highlight failure cases by showing that the algorithm can get stuck in spurious local minima.
****
Boosting the Throughput and Accelerator Utilization of Specialized CNN Inference Beyond Increasing Batch Size
Jack Kosaian, Amar Phanishayee, Matthai Philipose, Debadeepta Dey, Rashmi Vinayak
Datacenter vision systems widely use small, specialized convolutional neural networks (CNNs) trained on specific tasks for high-throughput inference. These settings employ accelerators with massive computational capacity, but which specialized CNNs underutilize due to having low arithmetic intensity. This results in suboptimal application-level throughput and poor returns on accelerator investment. Increasing batch size is the only known way to increase both application-level throughput and accelerator utilization for inference, but yields diminishing returns; specialized CNNs poorly utilize accelerators even with large batch size. We propose FoldedCNNs, a new approach to CNN design that increases inference throughput and utilization beyond large batch size. FoldedCNNs rethink the structure of inputs and layers of specialized CNNs to boost arithmetic intensity: in FoldedCNNs, f images with C channels each are concatenated into a single input with fC channels and jointly classified by a wider CNN. Increased arithmetic intensity in FoldedCNNs increases the throughput and GPU utilization of specialized CNN inference by up to 2.5x and 2.8x, with accuracy close to the original CNN in most cases.
****
NeRF-VAE: A Geometry Aware 3D Scene Generative Model
Adam R Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Sona Mokra, Danilo Jimenez Rezende
We propose NeRF-VAE, a 3D scene generative model that incorporates geometric structure via Neural Radiance Fields (NeRF) and differentiable volume rendering. In contrast to NeRF, our model takes into account shared structure across scenes, and is able to infer the structure of a novel scene—without the need to re-train—using amortized inference. NeRF-VAE's explicit 3D rendering process further contrasts previous generative models with convolution-based rendering which lacks geometric structure. Our model is a VAE that learns a distribution over radiance fields by conditioning them on a latent scene representation. We show that, once trained, NeRF-VAE is able to infer and render geometrically-consistent scenes from previously unseen 3D environments of synthetic scenes using very few input images. We further demonstrate that NeRF-VAE generalizes well to out-of-distribution cameras, while convolutional models do not. Finally, we introduce and study an attention-based conditioning mechanism of NeRF-VAE's decoder, which improves model performance.
****
Active Testing: Sample-Efficient Model Evaluation
Jannik Kossen, Sebastian Farquhar, Yarin Gal, Tom Rainforth
We introduce a new framework for sample-efficient model evaluation that we call active testing. While approaches like active learning reduce the number of labels needed for model training, existing literature largely ignores the cost of labeling test data, typically unrealistically assuming large test sets for model evaluation. This creates a disconnect to real applications, where test labels are important and just as expensive, e.g. for optimizing hyperparameters. Active testing addresses this by carefully selecting the test points to label, ensuring model evaluation is sample-efficient. To this end, we derive theoretically-grounded and intuitive acquisition strategies that are specifically tailored to the goa

ls of active testing, noting these are distinct to those of active learning. As actively selecting labels introduces a bias; we further show how to remove this bias while reducing the variance of the estimator at the same time. Active testi ng is easy to implement and can be applied to any supervised machine learning me thod. We demonstrate its effectiveness on models including WideResNets and Gauss ian processes on datasets including Fashion-MNIST and CIFAR-100.
****
High Confidence Generalization for Reinforcement Learning
James Kostas, Yash Chandak, Scott M Jordan, Georgios Theocharous, Philip Thomas
We present several classes of reinforcement learning algorithms that safely gene ralize to Markov decision processes (MDPs) not seen during training. Specificall y, we study the setting in which some set of MDPs is accessible for training. Th e goal is to generalize safely to MDPs that are sampled from the same distributi on, but which may not be in the set accessible for training. For various definit ions of safety, our algorithms give probabilistic guarantees that agents can saf ely generalize to MDPs that are sampled from the same distribution but are not n ecessarily in the training set. These algorithms are a type of Seldonian algorit hm (Thomas et al., 2019), which is a class of machine learning algorithms that r eturn models with probabilistic safety guarantees for user-specified definitions of safety.
****
Offline Reinforcement Learning with Fisher Divergence Critic Regularization
Ilya Kostrikov, Rob Fergus, Jonathan Tompson, Ofir Nachum
Many modern approaches to offline Reinforcement Learning (RL) utilize behavior r egularization, typically augmenting a model-free actor critic algorithm with a p enalty measuring divergence of the policy from the offline data. In this work, w e propose an alternative approach to encouraging the learned policy to stay clos e to the data, namely parameterizing the critic as the log-behavior-policy, whic h generated the offline data, plus a state-action value offset term, which can b e learned using a neural network. Behavior regularization then corresponds to an appropriate regularizer on the offset term. We propose using a gradient penalty regularizer for the offset term and demonstrate its equivalence to Fisher diver gence regularization, suggesting connections to the score matching and generativ e energy-based model literature. We thus term our resulting algorithm Fisher-BRC (Behavior Regularized Critic). On standard offline RL benchmarks, Fisher-BRC ac hieves both improved performance and faster convergence over existing state-of-t he-art methods.
****
ADOM: Accelerated Decentralized Optimization Method for Time-Varying Networks
Dmitry Kovalev, Egor Shulgin, Peter Richtarik, Alexander V Rogozin, Alexander Ga snikov
We propose ADOM – an accelerated method for smooth and strongly convex decentral ized optimization over time-varying networks. ADOM uses a dual oracle, i.e., we assume access to the gradient of the Fenchel conjugate of the individual loss fu nctions. Up to a constant factor, which depends on the network structure only, i ts communication complexity is the same as that of accelerated Nesterov gradient method. To the best of our knowledge, only the algorithm of Rogozin et al. (201 9) has a convergence rate with similar properties. However, their algorithm conv erges under the very restrictive assumption that the number of network changes c an not be greater than a tiny percentage of the number of iterations. This assum ption is hard to satisfy in practice, as the network topology changes usually ca n not be controlled. In contrast, ADOM merely requires the network to stay conne cted throughout time.
****
Revisiting Peng’s Q($λ$) for Modern Reinforcement Learning
Tadashi Kozuno, Yunhao Tang, Mark Rowland, Remi Munos, Steven Kapturowski, Will Dabney, Michal Valko, David Abel
Off-policy multi-step reinforcement learning algorithms consist of conservative and non-conservative algorithms: the former actively cut traces, whereas the lat ter do not. Recently, Munos et al. (2016) proved the convergence of conservative

algorithms to an optimal Q-function. In contrast, non-conservative algorithms are thought to be unsafe and have a limited or no theoretical guarantee. Nonetheless, recent studies have shown that non-conservative algorithms empirically outperform conservative ones. Motivated by the empirical results and the lack of theory, we carry out theoretical analyses of Peng's Q(λ), a representative example of non-conservative algorithms. We prove that \emph{it also converges to an optimal policy} provided that the behavior policy slowly tracks a greedy policy in a way similar to conservative policy iteration. Such a result has been conjectured to be true but has not been proven. We also experiment with Peng's Q(λ) in complex continuous control tasks, confirming that Peng's Q(λ) often outperforms conservative algorithms despite its simplicity. These results indicate that Peng's Q(λ), which was thought to be unsafe, is a theoretically-sound and practically effective algorithm.
****
Adapting to misspecification in contextual bandits with offline regression oracles
Sanath Kumar Krishnamurthy, Vitor Hadad, Susan Athey
Computationally efficient contextual bandits are often based on estimating a predictive model of rewards given contexts and arms using past data. However, when the reward model is not well-specified, the bandit algorithm may incur unexpected regret, so recent work has focused on algorithms that are robust to misspecification. We propose a simple family of contextual bandit algorithms that adapt to misspecification error by reverting to a good safe policy when there is evidence that misspecification is causing a regret increase. Our algorithm requires only an offline regression oracle to ensure regret guarantees that gracefully degrade in terms of a measure of the average misspecification level. Compared to prior work, we attain similar regret guarantees, but we do no rely on a master algorithm, and do not require more robust oracles like online or constrained regression oracles (e.g., Foster et al. (2020), Krishnamurthy et al. (2020)). This allows us to design algorithms for more general function approximation classes.
****
Out-of-Distribution Generalization via Risk Extrapolation (REx)
David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, Aaron Courville
Distributional shift is one of the major obstacles when transferring machine learning prediction systems from the lab to the real world. To tackle this problem, we assume that variation across training domains is representative of the variation we might encounter at test time, but also that shifts at test time may be more extreme in magnitude. In particular, we show that reducing differences in risk across training domains can reduce a model's sensitivity to a wide range of extreme distributional shifts, including the challenging setting where the input contains both causal and anti-causal elements. We motivate this approach, Risk Extrapolation (REx), as a form of robust optimization over a perturbation set of extrapolated domains (MM-REx), and propose a penalty on the variance of training risks (V-REx) as a simpler variant. We prove that variants of REx can recover the causal mechanisms of the targets, while also providing robustness to changes in the input distribution ("covariate shift"). By appropriately trading-off robustness to causally induced distributional shifts and covariate shift, REx is able to outperform alternative methods such as Invariant Risk Minimization in situations where these types of shift co-occur.
****
Near-Optimal Confidence Sequences for Bounded Random Variables
Arun K Kuchibhotla, Qinqing Zheng
Many inference problems, such as sequential decision problems like A/B testing, adaptive sampling schemes like bandit selection, are often online in nature. The fundamental problem for online inference is to provide a sequence of confidence intervals that are valid uniformly over the growing-into-infinity sample sizes. To address this question, we provide a near-optimal confidence sequence for bounded random variables by utilizing Bentkus' concentration results. We show that it improves on the existing approaches that use the Cram{é}r-Chernoff technique

such as the Hoeffding, Bernstein, and Bennett inequalities. The resulting confid ence sequence is confirmed to be favorable in synthetic coverage problems, adapt ive stopping algorithms, and multi-armed bandit problems.
****
Differentially Private Bayesian Inference for Generalized Linear Models
Tejas Kulkarni, Joonas Jälkö, Antti Koskela, Samuel Kaski, Antti Honkela
Generalized linear models (GLMs) such as logistic regression are among the most widely used arms in data analyst’s repertoire and often used on sensitive datase ts. A large body of prior works that investigate GLMs under differential privacy (DP) constraints provide only private point estimates of the regression coeffic ients, and are not able to quantify parameter uncertainty. In this work, with lo gistic and Poisson regression as running examples, we introduce a generic noise-aware DP Bayesian inference method for a GLM at hand, given a noisy sum of summa ry statistics. Quantifying uncertainty allows us to determine which of the regre ssion coefficients are statistically significantly different from zero. We provi de a previously unknown tight privacy analysis and experimentally demonstrate th at the posteriors obtained from our model, while adhering to strong privacy guar antees, are close to the non-private posteriors.
****
Bayesian Structural Adaptation for Continual Learning
Abhishek Kumar, Sunabha Chatterjee, Piyush Rai
Continual Learning is a learning paradigm where learning systems are trained on a sequence of tasks. The goal here is to perform well on the current task withou t suffering from a performance drop on the previous tasks. Two notable direction s among the recent advances in continual learning with neural networks are (1) v ariational Bayes based regularization by learning priors from previous tasks, an d, (2) learning the structure of deep networks to adapt to new tasks. So far, th ese two approaches have been largely orthogonal. We present a novel Bayesian fra mework based on continually learning the structure of deep neural networks, to u nify these distinct yet complementary approaches. The proposed framework learns the deep structure for each task by learning which weights to be used, and suppo rts inter-task transfer through the overlapping of different sparse subsets of w eights learned by different tasks. An appealing aspect of our proposed continual learning framework is that it is applicable to both discriminative (supervised) and generative (unsupervised) settings. Experimental results on supervised and unsupervised benchmarks demonstrate that our approach performs comparably or bet ter than recent advances in continual learning.
****
Implicit rate-constrained optimization of non-decomposable objectives
Abhishek Kumar, Harikrishna Narasimhan, Andrew Cotter
We consider a popular family of constrained optimization problems arising in mac hine learning that involve optimizing a non-decomposable evaluation metric with a certain thresholded form, while constraining another metric of interest. Examp les of such problems include optimizing false negative rate at a fixed false pos itive rate, optimizing precision at a fixed recall, optimizing the area under th e precision-recall or ROC curves, etc. Our key idea is to formulate a rate-const rained optimization that expresses the threshold parameter as a function of the model parameters via the Implicit Function theorem. We show how the resulting op timization problem can be solved using standard gradient based methods. Experime nts on benchmark datasets demonstrate the effectiveness of our proposed method o ver existing state-of-the-art approaches for these problems.
****
A Scalable Second Order Method for Ill-Conditioned Matrix Completion from Few Sa mples
Christian Kümmerle, Claudio M. Verdun
We propose an iterative algorithm for low-rank matrix completion with that can b e interpreted as an iteratively reweighted least squares (IRLS) algorithm, a sad dle-escaping smoothing Newton method or a variable metric proximal gradient meth od applied to a non-convex rank surrogate. It combines the favorable data-effici ency of previous IRLS approaches with an improved scalability by several orders

of magnitude. We establish the first local convergence guarantee from a minimal number of samples for that class of algorithms, showing that the method attains a local quadratic convergence rate. Furthermore, we show that the linear systems to be solved are well-conditioned even for very ill-conditioned ground truth matrices. We provide extensive experiments, indicating that unlike many state-of-the-art approaches, our method is able to complete very ill-conditioned matrices with a condition number of up to 10^{10} from few samples, while being competitive in its scalability.
****
Meta-Thompson Sampling
Branislav Kveton, Mikhail Konobeev, Manzil Zaheer, Chih-Wei Hsu, Martin Mladenov, Craig Boutilier, Csaba Szepesvari
Efficient exploration in bandits is a fundamental online learning problem. We propose a variant of Thompson sampling that learns to explore better as it interacts with bandit instances drawn from an unknown prior. The algorithm meta-learns the prior and thus we call it MetaTS. We propose several efficient implementations of MetaTS and analyze it in Gaussian bandits. Our analysis shows the benefit of meta-learning and is of a broader interest, because we derive a novel prior-dependent Bayes regret bound for Thompson sampling. Our theory is complemented by empirical evaluation, which shows that MetaTS quickly adapts to the unknown prior.
****
Targeted Data Acquisition for Evolving Negotiation Agents
Minae Kwon, Siddharth Karamcheti, Mariano-Florentino Cuellar, Dorsa Sadigh
Successful negotiators must learn how to balance optimizing for self-interest and cooperation. Yet current artificial negotiation agents often heavily depend on the quality of the static datasets they were trained on, limiting their capacity to fashion an adaptive response balancing self-interest and cooperation. For this reason, we find that these agents can achieve either high utility or cooperation, but not both. To address this, we introduce a targeted data acquisition framework where we guide the exploration of a reinforcement learning agent using annotations from an expert oracle. The guided exploration incentivizes the learning agent to go beyond its static dataset and develop new negotiation strategies. We show that this enables our agents to obtain higher-reward and more Pareto-optimal solutions when negotiating with both simulated and human partners compared to standard supervised learning and reinforcement learning methods. This trend additionally holds when comparing agents using our targeted data acquisition framework to variants of agents trained with a mix of supervised learning and reinforcement learning, or to agents using tailored reward functions that explicitly optimize for utility and Pareto-optimality.
****
ASAM: Adaptive Sharpness-Aware Minimization for Scale-Invariant Learning of Deep Neural Networks
Jungmin Kwon, Jeongseop Kim, Hyunseo Park, In Kwon Choi
Recently, learning algorithms motivated from sharpness of loss surface as an effective measure of generalization gap have shown state-of-the-art performances. Nevertheless, sharpness defined in a rigid region with a fixed radius, has a drawback in sensitivity to parameter re-scaling which leaves the loss unaffected, leading to weakening of the connection between sharpness and generalization gap. In this paper, we introduce the concept of adaptive sharpness which is scale-invariant and propose the corresponding generalization bound. We suggest a novel learning method, adaptive sharpness-aware minimization (ASAM), utilizing the proposed generalization bound. Experimental results in various benchmark datasets show that ASAM contributes to significant improvement of model generalization performance.
****
On the price of explainability for some clustering problems
Eduardo S Laber, Lucas Murtinho
The price of explainability for a clustering task can be defined as the unavoidable loss, in terms of the objective function, if we force the final partition to

be explainable. Here, we study this price for the following clustering problems : k-means, k-medians, k-centers and maximum-spacing. We provide upper and lower bounds for a natural model where explainability is achieved via decision trees. For the k-means and k-medians problems our upper bounds improve those obtained by [Dasgupta et. al, ICML 20] for low dimensions. Another contribution is a simple and efficient algorithm for building explainable clusterings for the k-means problem. We provide empirical evidence that its performance is better than the current state of the art for decision-tree based explainable clustering.
****
Adaptive Newton Sketch: Linear-time Optimization with Quadratic Convergence and Effective Hessian Dimensionality
Jonathan Lacotte, Yifei Wang, Mert Pilanci
We propose a randomized algorithm with quadratic convergence rate for convex optimization problems with a self-concordant, composite, strongly convex objective function. Our method is based on performing an approximate Newton step using a random projection of the Hessian. Our first contribution is to show that, at each iteration, the embedding dimension (or sketch size) can be as small as the effective dimension of the Hessian matrix. Leveraging this novel fundamental result, we design an algorithm with a sketch size proportional to the effective dimension and which exhibits a quadratic rate of convergence. This result dramatically improves on the classical linear-quadratic convergence rates of state-of-the-art sub-sampled Newton methods. However, in most practical cases, the effective dimension is not known beforehand, and this raises the question of how to pick a sketch size as small as the effective dimension while preserving a quadratic convergence rate. Our second and main contribution is thus to propose an adaptive sketch size algorithm with quadratic convergence rate and which does not require prior knowledge or estimation of the effective dimension: at each iteration, it starts with a small sketch size, and increases it until quadratic progress is achieved. Importantly, we show that the embedding dimension remains proportional to the effective dimension throughout the entire path and that our method achieves state-of-the-art computational complexity for solving convex optimization programs with a strongly convex component. We discuss and illustrate applications to linear and quadratic programming, as well as logistic regression and other generalized linear models.
****
Generalization Bounds in the Presence of Outliers: a Median-of-Means Study
Pierre Laforgue, Guillaume Staerman, Stephan Clémençon
In contrast to the empirical mean, the Median-of-Means (MoM) is an estimator of the mean θ of a square integrable r.v. Z, around which accurate nonasymptotic confidence bounds can be built, even when Z does not exhibit a sub-Gaussian tail behavior. Thanks to the high confidence it achieves on heavy-tailed data, MoM has found various applications in machine learning, where it is used to design training procedures that are not sensitive to atypical observations. More recently, a new line of work is now trying to characterize and leverage MoM's ability to deal with corrupted data. In this context, the present work proposes a general study of MoM's concentration properties under the contamination regime, that provides a clear understanding on the impact of the outlier proportion and the number of blocks chosen. The analysis is extended to (multisample) U-statistics, i.e. averages over tuples of observations, that raise additional challenges due to the dependence induced. Finally, we show that the latter bounds can be used in a straightforward fashion to derive generalization guarantees for pairwise learning in a contaminated setting, and propose an algorithm to compute provably reliable decision functions.
****
Model Fusion for Personalized Learning
Thanh Chi Lam, Nghia Hoang, Bryan Kian Hsiang Low, Patrick Jaillet
Production systems operating on a growing domain of analytic services often require generating warm-start solution models for emerging tasks with limited data. One potential approach to address this warm-start challenge is to adopt meta lea

rning to generate a base model that can be adapted to solve unseen tasks with minimal fine-tuning. This however requires the training processes of previous solution models of existing tasks to be synchronized. This is not possible if these models were pre-trained separately on private data owned by different entities and cannot be synchronously re-trained. To accommodate for such scenarios, we develop a new personalized learning framework that synthesizes customized models for unseen tasks via fusion of independently pre-trained models of related tasks. We establish performance guarantee for the proposed framework and demonstrate its effectiveness on both synthetic and real datasets.
****
Gradient Disaggregation: Breaking Privacy in Federated Learning by Reconstructing the User Participant Matrix
Maximilian Lam, Gu-Yeon Wei, David Brooks, Vijay Janapa Reddi, Michael Mitzenmacher
We show that aggregated model updates in federated learning may be insecure. An untrusted central server may disaggregate user updates from sums of updates across participants given repeated observations, enabling the server to recover privileged information about individual users' private training data via traditional gradient inference attacks. Our method revolves around reconstructing participant information (e.g: which rounds of training users participated in) from aggregated model updates by leveraging summary information from device analytics commonly used to monitor, debug, and manage federated learning systems. Our attack is parallelizable and we successfully disaggregate user updates on settings with up to thousands of participants. We quantitatively and qualitatively demonstrate significant improvements in the capability of various inference attacks on the disaggregated updates. Our attack enables the attribution of learned properties to individual users, violating anonymity, and shows that a determined central server may undermine the secure aggregation protocol to break individual users' data privacy in federated learning.
****
Stochastic Multi-Armed Bandits with Unrestricted Delay Distributions
Tal Lancewicki, Shahar Segal, Tomer Koren, Yishay Mansour
We study the stochastic Multi-Armed Bandit (MAB) problem with random delays in the feedback received by the algorithm. We consider two settings: the {\it reward dependent} delay setting, where realized delays may depend on the stochastic rewards, and the {\it reward-independent} delay setting. Our main contribution is algorithms that achieve near-optimal regret in each of the settings, with an additional additive dependence on the quantiles of the delay distribution. Our results do not make any assumptions on the delay distributions: in particular, we do not assume they come from any parametric family of distributions and allow for unbounded support and expectation; we further allow for the case of infinite delays where the algorithm might occasionally not observe any feedback.
****
Discovering symbolic policies with deep reinforcement learning
Mikel Landajuela, Brenden K Petersen, Sookyung Kim, Claudio P Santiago, Ruben Glatt, Nathan Mundhenk, Jacob F Pettit, Daniel Faissol
Deep reinforcement learning (DRL) has proven successful for many difficult control problems by learning policies represented by neural networks. However, the complexity of neural network-based policies{—}involving thousands of composed non-linear operators{—}can render them problematic to understand, trust, and deploy. In contrast, simple policies comprising short symbolic expressions can facilitate human understanding, while also being transparent and exhibiting predictable behavior. To this end, we propose deep symbolic policy, a novel approach to directly search the space of symbolic policies. We use an autoregressive recurrent neural network to generate control policies represented by tractable mathematical expressions, employing a risk-seeking policy gradient to maximize performance of the generated policies. To scale to environments with multi-dimensional action spaces, we propose an "anchoring" algorithm that distills pre-trained neural network-based policies into fully symbolic policies, one action dimension at a time. We also introduce two novel methods to improve exploration in DRL-based combi

natorial optimization, building on ideas of entropy regularization and distribution initialization. Despite their dramatically reduced complexity, we demonstrate that discovered symbolic policies outperform seven state-of-the-art DRL algorithms in terms of average rank and average normalized episodic reward across eight benchmark environments.
****
Graph Cuts Always Find a Global Optimum for Potts Models (With a Catch)
Hunter Lang, David Sontag, Aravindan Vijayaraghavan
We prove that the alpha-expansion algorithm for MAP inference always returns a globally optimal assignment for Markov Random Fields with Potts pairwise potentials, with a catch: the returned assignment is only guaranteed to be optimal for an instance within a small perturbation of the original problem instance. In other words, all local minima with respect to expansion moves are global minima to slightly perturbed versions of the problem. On "real-world" instances, MAP assignments of small perturbations of the problem should be very similar to the MAP assignment(s) of the original problem instance. We design an algorithm that can certify whether this is the case in practice. On several MAP inference problem instances from computer vision, this algorithm certifies that MAP solutions to all of these perturbations are very close to solutions of the original instance. These results taken together give a cohesive explanation for the good performance of "graph cuts" algorithms in practice. Every local expansion minimum is a global minimum in a small perturbation of the problem, and all of these global minima are close to the original solution.
****
Efficient Message Passing for 0–1 ILPs with Binary Decision Diagrams
Jan-Hendrik Lange, Paul Swoboda
We present a message passing method for 0{–}1 integer linear programs. Our algorithm is based on a decomposition of the original problem into subproblems that are represented as binary deci- sion diagrams. The resulting Lagrangean dual is solved iteratively by a series of efficient block coordinate ascent steps. Our method has linear iteration complexity in the size of the decomposi- tion and can be effectively parallelized. The char- acteristics of our approach are desirable towards solving ever larger problems arising in structured prediction. We present experimental results on combinatorial problems from MAP inference for Markov Random Fields, quadratic assignment, discrete tomography and cell tracking for develop- mental biology and show promising performance.
****
CountSketches, Feature Hashing and the Median of Three
Kasper Green Larsen, Rasmus Pagh, Jakub T■tek
In this paper, we revisit the classic CountSketch method, which is a sparse, random projection that transforms a (high-dimensional) Euclidean vector v to a vector of dimension $(2t-1) s$, where $t, s > 0$ are integer parameters. It is known that a CountSketch allows estimating coordinates of v with variance bounded by $\|v\|_2^2/s$. For $t > 1$, the estimator takes the median of $2t-1$ independent estimates, and the probability that the estimate is off by more than $2 \|v\|_2/\sqrt{s}$ is exponentially small in t. This suggests choosing t to be logarithmic in a desired inverse failure probability. However, implementations of CountSketch often use a small, constant t. Previous work only predicts a constant factor improvement in this setting. Our main contribution is a new analysis of CountSketch, showing an improvement in variance to $O(\min\{\|v\|_1^2/s^2,\|v\|_2^2/s\})$ when $t > 1$. That is, the variance decreases proportionally to s^{-2}, asymptotically for large enough s.
****
MorphVAE: Generating Neural Morphologies from 3D-Walks using a Variational Autoencoder with Spherical Latent Space
Sophie C. Laturnus, Philipp Berens
For the past century, the anatomy of a neuron has been considered one of its defining features: The shape of a neuron’s dendrites and axon fundamentally determines what other neurons it can connect to. These neurites have been described using mathematical tools e.g. in the context of cell type classification, but gener

ative models of these structures have only rarely been proposed and are often computationally inefficient. Here we propose MorphVAE, a sequence-to-sequence variational autoencoder with spherical latent space as a generative model for neural morphologies. The model operates on walks within the tree structure of a neuron and can incorporate expert annotations on a subset of the data using semi-supervised learning. We develop our model on artificially generated toy data and evaluate its performance on dendrites of excitatory cells and axons of inhibitory cells of mouse motor cortex (M1) and dendrites of retinal ganglion cells. We show that the learned latent feature space allows for better cell type discrimination than other commonly used features. By sampling new walks from the latent space we can easily construct new morphologies with a specified degree of similarity to their reference neuron, providing an efficient generative model for neural morphologies.
****
Improved Regret Bound and Experience Replay in Regularized Policy Iteration
Nevena Lazic, Dong Yin, Yasin Abbasi-Yadkori, Csaba Szepesvari
In this work, we study algorithms for learning in infinite-horizon undiscounted Markov decision processes (MDPs) with function approximation. We first show that the regret analysis of the Politex algorithm (a version of regularized policy iteration) can be sharpened from $O(T^{3/4})$ to $O(\sqrt{T})$ under nearly identical assumptions, and instantiate the bound with linear function approximation. Our result provides the first high-probability $O(\sqrt{T})$ regret bound for a computationally efficient algorithm in this setting. The exact implementation of Politex with neural network function approximation is inefficient in terms of memory and computation. Since our analysis suggests that we need to approximate the average of the action-value functions of past policies well, we propose a simple efficient implementation where we train a single Q-function on a replay buffer with past data. We show that this often leads to superior performance over other implementation choices, especially in terms of wall-clock time. Our work also provides a novel theoretical justification for using experience replay within policy iteration algorithms.
****
LAMDA: Label Matching Deep Domain Adaptation
Trung Le, Tuan Nguyen, Nhat Ho, Hung Bui, Dinh Phung
Deep domain adaptation (DDA) approaches have recently been shown to perform better than their shallow rivals with better modeling capacity on complex domains (e.g., image, structural data, and sequential data). The underlying idea is to learn domain invariant representations on a latent space that can bridge the gap between source and target domains. Several theoretical studies have established insightful understanding and the benefit of learning domain invariant features; however, they are usually limited to the case where there is no label shift, hence hindering its applicability. In this paper, we propose and study a new challenging setting that allows us to use a Wasserstein distance (WS) to not only quantify the data shift but also to define the label shift directly. We further develop a theory to demonstrate that minimizing the WS of the data shift leads to closing the gap between the source and target data distributions on the latent space (e.g., an intermediate layer of a deep net), while still being able to quantify the label shift with respect to this latent space. Interestingly, our theory can consequently explain certain drawbacks of learning domain invariant features on the latent space. Finally, grounded on the results and guidance of our developed theory, we propose the Label Matching Deep Domain Adaptation (LAMDA) approach that outperforms baselines on real-world datasets for DA problems.
****
Gaussian Process-Based Real-Time Learning for Safety Critical Applications
Armin Lederer, Alejandro J Ordóñez Conejo, Korbinian A Maier, Wenxin Xiao, Jonas Umlauft, Sandra Hirche
The safe operation of physical systems typically relies on high-quality models. Since a continuous stream of data is generated during run-time, such models are often obtained through the application of Gaussian process regression because it provides guarantees on the prediction error. Due to its high computational comp

lexity, Gaussian process regression must be used offline on batches of data, whi ch prevents applications, where a fast adaptation through online learning is nec essary to ensure safety. In order to overcome this issue, we propose the LoG-GP. It achieves a logarithmic update and prediction complexity in the number of tra ining points through the aggregation of locally active Gaussian process models. Under weak assumptions on the aggregation scheme, it inherits safety guarantees from exact Gaussian process regression. These theoretical advantages are exempla rily exploited in the design of a safe and data-efficient, online-learning contr ol policy. The efficiency and performance of the proposed real-time learning app roach is demonstrated in a comparison to state-of-the-art methods.
****
Sharing Less is More: Lifelong Learning in Deep Networks with Selective Layer Tr ansfer
Seungwon Lee, Sima Behpour, Eric Eaton
Effective lifelong learning across diverse tasks requires the transfer of divers e knowledge, yet transferring irrelevant knowledge may lead to interference and catastrophic forgetting. In deep networks, transferring the appropriate granular ity of knowledge is as important as the transfer mechanism, and must be driven b y the relationships among tasks. We first show that the lifelong learning perfor mance of several current deep learning architectures can be significantly improv ed by transfer at the appropriate layers. We then develop an expectation-maximiz ation (EM) method to automatically select the appropriate transfer configuration and optimize the task network weights. This EM-based selective transfer is high ly effective, balancing transfer performance on all tasks with avoiding catastro phic forgetting, as demonstrated on three algorithms in several lifelong object classification scenarios.
****
Fair Selective Classification Via Sufficiency
Joshua K Lee, Yuheng Bu, Deepta Rajan, Prasanna Sattigeri, Rameswar Panda, Subhr o Das, Gregory W Wornell
Selective classification is a powerful tool for decision-making in scenarios whe re mistakes are costly but abstentions are allowed. In general, by allowing a cl assifier to abstain, one can improve the performance of a model at the cost of r educing coverage and classifying fewer samples. However, recent work has shown, in some cases, that selective classification can magnify disparities between gro ups, and has illustrated this phenomenon on multiple real-world datasets. We pro ve that the sufficiency criterion can be used to mitigate these disparities by e nsuring that selective classification increases performance on all groups, and i ntroduce a method for mitigating the disparity in precision across the entire co verage scale based on this criterion. We then provide an upper bound on the cond itional mutual information between the class label and sensitive attribute, cond itioned on the learned features, which can be used as a regularizer to achieve f airer selective classification. The effectiveness of the method is demonstrated on the Adult, CelebA, Civil Comments, and CheXpert datasets.
****
On-the-fly Rectification for Robust Large-Vocabulary Topic Inference
Moontae Lee, Sungjun Cho, Kun Dong, David Mimno, David Bindel
Across many data domains, co-occurrence statistics about the joint appearance of objects are powerfully informative. By transforming unsupervised learning probl ems into decompositions of co-occurrence statistics, spectral algorithms provide transparent and efficient algorithms for posterior inference such as latent top ic analysis and community detection. As object vocabularies grow, however, it be comes rapidly more expensive to store and run inference algorithms on co-occurre nce statistics. Rectifying co-occurrence, the key process to uphold model assump tions, becomes increasingly more vital in the presence of rare terms, but curren t techniques cannot scale to large vocabularies. We propose novel methods that s imultaneously compress and rectify co-occurrence statistics, scaling gracefully with the size of vocabulary and the dimension of latent space. We also present n ew algorithms learning latent variables from the compressed statistics, and veri fy that our methods perform comparably to previous approaches on both textual an

d non-textual data.
****
Unsupervised Embedding Adaptation via Early-Stage Feature Reconstruction for Few-Shot Classification
Dong Hoon Lee, Sae-Young Chung
We propose unsupervised embedding adaptation for the downstream few-shot classification task. Based on findings that deep neural networks learn to generalize before memorizing, we develop Early-Stage Feature Reconstruction (ESFR) — a novel adaptation scheme with feature reconstruction and dimensionality-driven early stopping that finds generalizable features. Incorporating ESFR consistently improves the performance of baseline methods on all standard settings, including the recently proposed transductive method. ESFR used in conjunction with the transductive method further achieves state-of-the-art performance on mini-ImageNet, tiered-ImageNet, and CUB; especially with 1.2% 2.0% improvements in accuracy over the previous best performing method on 1-shot setting.
****
Continual Learning in the Teacher-Student Setup: Impact of Task Similarity
Sebastian Lee, Sebastian Goldt, Andrew Saxe
Continual learning{—}the ability to learn many tasks in sequence{—}is critical for artificial learning systems. Yet standard training methods for deep networks often suffer from catastrophic forgetting, where learning new tasks erases knowledge of the earlier tasks. While catastrophic forgetting labels the problem, the theoretical reasons for interference between tasks remain unclear. Here, we attempt to narrow this gap between theory and practice by studying continual learning in the teacher-student setup. We extend previous analytical work on two-layer networks in the teacher-student setup to multiple teachers. Using each teacher to represent a different task, we investigate how the relationship between teachers affects the amount of forgetting and transfer exhibited by the student when the task switches. In line with recent work, we find that when tasks depend on similar features, intermediate task similarity leads to greatest forgetting. However, feature similarity is only one way in which tasks may be related. The teacher-student approach allows us to disentangle task similarity at the level of \emph{readouts} (hidden-to-output weights) as well as \emph{features} (input-to-hidden weights). We find a complex interplay between both types of similarity, initial transfer/forgetting rates, maximum transfer/forgetting, and the long-time (post-switch) amount of transfer/forgetting. Together, these results help illuminate the diverse factors contributing to catastrophic forgetting.
****
OptiDICE: Offline Policy Optimization via Stationary Distribution Correction Estimation
Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, Kee-Eung Kim
We consider the offline reinforcement learning (RL) setting where the agent aims to optimize the policy solely from the data without further environment interactions. In offline RL, the distributional shift becomes the primary source of difficulty, which arises from the deviation of the target policy being optimized from the behavior policy used for data collection. This typically causes overestimation of action values, which poses severe problems for model-free algorithms that use bootstrapping. To mitigate the problem, prior offline RL algorithms often used sophisticated techniques that encourage underestimation of action values, which introduces an additional set of hyperparameters that need to be tuned properly. In this paper, we present an offline RL algorithm that prevents overestimation in a more principled way. Our algorithm, OptiDICE, directly estimates the stationary distribution corrections of the optimal policy and does not rely on policy-gradients, unlike previous offline RL algorithms. Using an extensive set of benchmark datasets for offline RL, we show that OptiDICE performs competitively with the state-of-the-art methods.
****
SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning
Kimin Lee, Michael Laskin, Aravind Srinivas, Pieter Abbeel

Off-policy deep reinforcement learning (RL) has been successful in a range of ch
allenging domains. However, standard off-policy RL algorithms can suffer from se
veral issues, such as instability in Q-learning and balancing exploration and ex
ploitation. To mitigate these issues, we present SUNRISE, a simple unified ensem
ble method, which is compatible with various off-policy RL algorithms. SUNRISE i
ntegrates two key ingredients: (a) ensemble-based weighted Bellman backups, whic
h re-weight target Q-values based on uncertainty estimates from a Q-ensemble, an
d (b) an inference method that selects actions using the highest upper-confidenc
e bounds for efficient exploration. By enforcing the diversity between agents us
ing Bootstrap with random initialization, we show that these different ideas are
 largely orthogonal and can be fruitfully integrated, together further improving
 the performance of existing off-policy RL algorithms, such as Soft Actor-Critic
 and Rainbow DQN, for both continuous and discrete control tasks on both low-dim
ensional and high-dimensional environments.
****
Achieving Near Instance-Optimality and Minimax-Optimality in Stochastic and Adve
rsarial Linear Bandits Simultaneously
Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, Xiaojin Zhang
In this work, we develop linear bandit algorithms that automatically adapt to di
fferent environments. By plugging a novel loss estimator into the optimization p
roblem that characterizes the instance-optimal strategy, our first algorithm not
 only achieves nearly instance-optimal regret in stochastic environments, but al
so works in corrupted environments with additional regret being the amount of co
rruption, while the state-of-the-art (Li et al., 2019) achieves neither instance
-optimality nor the optimal dependence on the corruption amount. Moreover, by eq
uipping this algorithm with an adversarial component and carefully-designed test
ings, our second algorithm additionally enjoys minimax-optimal regret in complet
ely adversarial environments, which is the first of this kind to our knowledge.
Finally, all our guarantees hold with high probability, while existing instance-
optimal guarantees only hold in expectation.
****
PEBBLE: Feedback-Efficient Interactive Reinforcement Learning via Relabeling Exp
erience and Unsupervised Pre-training
Kimin Lee, Laura M Smith, Pieter Abbeel
Conveying complex objectives to reinforcement learning (RL) agents can often be
difficult, involving meticulous design of reward functions that are sufficiently
 informative yet easy enough to provide. Human-in-the-loop RL methods allow prac
titioners to instead interactively teach agents through tailored feedback; howev
er, such approaches have been challenging to scale since human feedback is very
expensive. In this work, we aim to make this process more sample- and feedback-e
fficient. We present an off-policy, interactive RL algorithm that capitalizes on
 the strengths of both feedback and off-policy learning. Specifically, we learn
a reward model by actively querying a teacher’s preferences between two clips of
 behavior and use it to train an agent. To enable off-policy learning, we relabe
l all the agent’s past experience when its reward model changes. We additionally
 show that pre-training our agents with unsupervised exploration substantially i
ncreases the mileage of its queries. We demonstrate that our approach is capable
 of learning tasks of higher complexity than previously considered by human-in-t
he-loop methods, including a variety of locomotion and robotic manipulation skil
ls. We also show that our method is able to utilize real-time human feedback to
effectively prevent reward exploitation and learn new behaviors that are difficu
lt to specify with standard reward functions.
****
Near-Optimal Linear Regression under Distribution Shift
Qi Lei, Wei Hu, Jason Lee
Transfer learning is essential when sufficient data comes from the source domain
, with scarce labeled data from the target domain. We develop estimators that ac
hieve minimax linear risk for linear regression problems under distribution shif
t. Our algorithms cover different transfer learning settings including covariate
 shift and model shift. We also consider when data are generated from either lin

ear or general nonlinear models. We show that linear minimax estimators are within an absolute constant of the minimax risk even among nonlinear estimators for various source/target distributions.
****

Stability and Generalization of Stochastic Gradient Methods for Minimax Problems
Yunwen Lei, Zhenhuan Yang, Tianbao Yang, Yiming Ying
Many machine learning problems can be formulated as minimax problems such as Generative Adversarial Networks (GANs), AUC maximization and robust estimation, to mention but a few. A substantial amount of studies are devoted to studying the convergence behavior of their stochastic gradient-type algorithms. In contrast, there is relatively little work on understanding their generalization, i.e., how the learning models built from training examples would behave on test examples. In this paper, we provide a comprehensive generalization analysis of stochastic gradient methods for minimax problems under both convex-concave and nonconvex-nonconcave cases through the lens of algorithmic stability. We establish a quantitative connection between stability and several generalization measures both in expectation and with high probability. For the convex-concave setting, our stability analysis shows that stochastic gradient descent ascent attains optimal generalization bounds for both smooth and nonsmooth minimax problems. We also establish generalization bounds for both weakly-convex-weakly-concave and gradient-dominated problems. We report preliminary experimental results to verify our theory.
****

Scalable Evaluation of Multi-Agent Reinforcement Learning with Melting Pot
Joel Z Leibo, Edgar A Dueñez-Guzman, Alexander Vezhnevets, John P Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charlie Beattie, Igor Mordatch, Thore Graepel
Existing evaluation suites for multi-agent reinforcement learning (MARL) do not assess generalization to novel situations as their primary objective (unlike supervised learning benchmarks). Our contribution, Melting Pot, is a MARL evaluation suite that fills this gap and uses reinforcement learning to reduce the human labor required to create novel test scenarios. This works because one agent’s behavior constitutes (part of) another agent’s environment. To demonstrate scalability, we have created over 80 unique test scenarios covering a broad range of research topics such as social dilemmas, reciprocity, resource sharing, and task partitioning. We apply these test scenarios to standard MARL training algorithms, and demonstrate how Melting Pot reveals weaknesses not apparent from training performance alone.
****

Better Training using Weight-Constrained Stochastic Dynamics
Benedict Leimkuhler, Tiffany J Vlaar, Timothée Pouchon, Amos Storkey
We employ constraints to control the parameter space of deep neural networks throughout training. The use of customised, appropriately designed constraints can reduce the vanishing/exploding gradients problem, improve smoothness of classification boundaries, control weight magnitudes and stabilize deep neural networks, and thus enhance the robustness of training algorithms and the generalization capabilities of neural networks. We provide a general approach to efficiently incorporate constraints into a stochastic gradient Langevin framework, allowing enhanced exploration of the loss landscape. We also present specific examples of constrained training methods motivated by orthogonality preservation for weight matrices and explicit weight normalizations. Discretization schemes are provided both for the overdamped formulation of Langevin dynamics and the underdamped form, in which momenta further improve sampling efficiency. These optimisation schemes can be used directly, without needing to adapt neural network architecture design choices or to modify the objective with regularization terms, and see performance improvements in classification tasks.
****

Globally-Robust Neural Networks
Klas Leino, Zifan Wang, Matt Fredrikson
The threat of adversarial examples has motivated work on training certifiably robust neural networks to facilitate efficient verification of local robustness at

inference time. We formalize a notion of global robustness, which captures the operational properties of on-line local robustness certification while yielding a natural learning objective for robust training. We show that widely-used architectures can be easily adapted to this objective by incorporating efficient global Lipschitz bounds into the network, yielding certifiably-robust models by construction that achieve state-of-the-art verifiable accuracy. Notably, this approach requires significantly less time and memory than recent certifiable training methods, and leads to negligible costs when certifying points on-line; for example, our evaluation shows that it is possible to train a large robust Tiny-Imagenet model in a matter of hours. Our models effectively leverage inexpensive global Lipschitz bounds for real-time certification, despite prior suggestions that tighter local bounds are needed for good performance; we posit this is possible because our models are specifically trained to achieve tighter global bounds. Namely, we prove that the maximum achievable verifiable accuracy for a given dataset is not improved by using a local bound.
****
Learning to Price Against a Moving Target
Renato Paes Leme, Balasubramanian Sivan, Yifeng Teng, Pratik Worah
In the Learning to Price setting, a seller posts prices over time with the goal of maximizing revenue while learning the buyer’s valuation. This problem is very well understood when values are stationary (fixed or iid). Here we study the problem where the buyer’s value is a moving target, i.e., they change over time either by a stochastic process or adversarially with bounded variation. In either case, we provide matching upper and lower bounds on the optimal revenue loss. Since the target is moving, any information learned soon becomes out-dated, which forces the algorithms to keep switching between exploring and exploiting phases.
****
SigGPDE: Scaling Sparse Gaussian Processes on Sequential Data
Maud Lemercier, Cristopher Salvi, Thomas Cass, Edwin V. Bonilla, Theodoros Damoulas, Terry J Lyons
Making predictions and quantifying their uncertainty when the input data is sequential is a fundamental learning challenge, recently attracting increasing attention. We develop SigGPDE, a new scalable sparse variational inference framework for Gaussian Processes (GPs) on sequential data. Our contribution is twofold. First, we construct inducing variables underpinning the sparse approximation so that the resulting evidence lower bound (ELBO) does not require any matrix inversion. Second, we show that the gradients of the GP signature kernel are solutions of a hyperbolic partial differential equation (PDE). This theoretical insight allows us to build an efficient back-propagation algorithm to optimize the ELBO. We showcase the significant computational gains of SigGPDE compared to existing methods, while achieving state-of-the-art performance for classification tasks on large datasets of up to 1 million multivariate time series.
****
Strategic Classification Made Practical
Sagi Levanon, Nir Rosenfeld
Strategic classification regards the problem of learning in settings where users can strategically modify their features to improve outcomes. This setting applies broadly, and has received much recent attention. But despite its practical significance, work in this space has so far been predominantly theoretical. In this paper we present a learning framework for strategic classification that is practical. Our approach directly minimizes the “strategic” empirical risk, which we achieve by differentiating through the strategic response of users. This provides flexibility that allows us to extend beyond the original problem formulation and towards more realistic learning scenarios. A series of experiments demonstrates the effectiveness of our approach on various learning settings.
****
Improved, Deterministic Smoothing for L_1 Certified Robustness
Alexander J Levine, Soheil Feizi
Randomized smoothing is a general technique for computing sample-dependent robustness guarantees against adversarial attacks for deep classifiers. Prior works o

n randomized smoothing against L_1 adversarial attacks use additive smoothing no
ise and provide probabilistic robustness guarantees. In this work, we propose a
non-additive and deterministic smoothing method, Deterministic Smoothing with Sp
litting Noise (DSSN). To develop DSSN, we first develop SSN, a randomized method
 which involves generating each noisy smoothing sample by first randomly splitti
ng the input space and then returning a representation of the center of the subd
ivision occupied by the input sample. In contrast to uniform additive smoothing,
 the SSN certification does not require the random noise components used to be i
ndependent. Thus, smoothing can be done effectively in just one dimension and ca
n therefore be efficiently derandomized for quantized data (e.g., images). To th
e best of our knowledge, this is the first work to provide deterministic "random
ized smoothing" for a norm-based adversarial threat model while allowing for an
arbitrary classifier (i.e., a deep model) to be used as a base classifier and wi
thout requiring an exponential number of smoothing samples. On CIFAR-10 and Imag
eNet datasets, we provide substantially larger L_1 robustness certificates compa
red to prior works, establishing a new state-of-the-art. The determinism of our
method also leads to significantly faster certificate computation. Code is avail
able at: https://github.com/alevine0/smoothingSplittingNoise.
****
BASE Layers: Simplifying Training of Large, Sparse Models
Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, Luke Zettlemoyer
We introduce a new balanced assignment of experts (BASE) layer for large languag
e models that greatly simplifies existing high capacity sparse layers. Sparse la
yers can dramatically improve the efficiency of training and inference by routin
g each token to specialized expert modules that contain only a small fraction of
 the model parameters. However, it can be difficult to learn balanced routing fu
nctions that make full use of the available experts; existing approaches typical
ly use routing heuristics or auxiliary expert-balancing loss functions. In contr
ast, we formulate token-to-expert allocation as a linear assignment problem, all
owing an optimal assignment in which each expert receives an equal number of tok
ens. This optimal assignment scheme improves efficiency by guaranteeing balanced
 compute loads, and also simplifies training by not requiring any new hyperparam
eters or auxiliary losses. Code is publicly released.
****
Run-Sort-ReRun: Escaping Batch Size Limitations in Sliced Wasserstein Generative
 Models
Jose Lezama, Wei Chen, Qiang Qiu
When training an implicit generative model, ideally one would like the generator
 to reproduce all the different modes and subtleties of the target distribution.
 Naturally, when comparing two empirical distributions, the larger the sample po
pulation, the more these statistical nuances can be captured. However, existing
objective functions are computationally constrained in the amount of samples the
y can consider by the memory required to process a batch of samples. In this pap
er, we build upon recent progress in sliced Wasserstein distances, a family of d
ifferentiable metrics for distribution discrepancy based on the Optimal Transpor
t paradigm. We introduce a procedure to train these distances with virtually any
 batch size, allowing the discrepancy measure to capture richer statistics and b
etter approximating the distance between the underlying continuous distributions
. As an example, we demonstrate the matching of the distribution of Inception fe
atures with batches of tens of thousands of samples, achieving FID scores that o
utperform state-of-the-art implicit generative models.
****
PAGE: A Simple and Optimal Probabilistic Gradient Estimator for Nonconvex Optimi
zation
Zhize Li, Hongyan Bao, Xiangliang Zhang, Peter Richtarik
In this paper, we propose a novel stochastic gradient estimator—ProbAbilistic Gr
adient Estimator (PAGE)—for nonconvex optimization. PAGE is easy to implement as
 it is designed via a small adjustment to vanilla SGD: in each iteration, PAGE u
ses the vanilla minibatch SGD update with probability p_t or reuses the previo
us gradient with a small adjustment, at a much lower computational cost, with pr

obability $1-p_t$. We give a simple formula for the optimal choice of p_t. Mor
eover, we prove the first tight lower bound $\Omega(n+\frac{\sqrt{n}}{\epsilon^2
})$ for nonconvex finite-sum problems, which also leads to a tight lower bound $
\Omega(b+\frac{\sqrt{b}}{\epsilon^2})$ for nonconvex online problems, where $b:=
 \min\{\frac{\sigma^2}{\epsilon^2}, n\}$. Then, we show that PAGE obtains the op
timal convergence results $O(n+\frac{\sqrt{n}}{\epsilon^2})$ (finite-sum) and $O
(b+\frac{\sqrt{b}}{\epsilon^2})$ (online) matching our lower bounds for both non
convex finite-sum and online problems. Besides, we also show that for nonconvex
functions satisfying the Polyak-■{ojasiewicz} (PL) condition, PAGE can automatic
ally switch to a faster linear convergence rate $O(\cdot\log \frac{1}{\epsilon})
$. Finally, we conduct several deep learning experiments (e.g., LeNet, VGG, ResN
et) on real datasets in PyTorch showing that PAGE not only converges much faster
 than SGD in training but also achieves the higher test accuracy, validating the
 optimal theoretical results and confirming the practical superiority of PAGE.
****
Tightening the Dependence on Horizon in the Sample Complexity of Q-Learning
Gen Li, Changxiao Cai, Yuxin Chen, Yuantao Gu, Yuting Wei, Yuejie Chi
Q-learning, which seeks to learn the optimal Q-function of a Markov decision pro
cess (MDP) in a model-free fashion, lies at the heart of reinforcement learning.
 Focusing on the synchronous setting (such that independent samples for all stat
e-action pairs are queried via a generative model in each iteration), substantia
l progress has been made recently towards understanding the sample efficiency of
 Q-learning. To yield an entrywise ε-accurate estimate of the optima
l Q-function, state-of-the-art theory requires at least an order of $\frac{|S||A
|}{(1-\gamma)^5\varepsilon^{2}}$ samples in the infinite-horizon γ-discou
nted setting. In this work, we sharpen the sample complexity of synchronous Q-le
arning to the order of $\frac{|S||A|}{(1-\gamma)^4\varepsilon^2}$ (up to some lo
garithmic factor) for any $0
  Cite this Paper

 BibTeX

```
@InProceedings{pmlr-v139-li21b,
  title = ■ {Tightening the Dependence on Horizon in the Sample Complexity of Q-
Learning},
  author =       {Li, Gen and Cai, Changxiao and Chen, Yuxin and Gu, Yuantao and
 Wei, Yuting and Chi, Yuejie},
  booktitle = ■ {Proceedings of the 38th International Conference on Machine Lea
rning},
  pages = ■ {6296--6306},
  year = ■ {2021},
  editor = ■ {Meila, Marina and Zhang, Tong},
  volume = ■ {139},
  series = ■ {Proceedings of Machine Learning Research},
  month = ■ {18--24 Jul},
  publisher =    {PMLR},
  pdf = ■ {http://proceedings.mlr.press/v139/li21b/li21b.pdf},
  url = ■ {https://proceedings.mlr.press/v139/li21b.html},
  abstract = ■ {Q-learning, which seeks to learn the optimal Q-function of a Mar
kov decision process (MDP) in a model-free fashion, lies at the heart of reinfor
cement learning. Focusing on the synchronous setting (such that independent samp
les for all state-action pairs are queried via a generative model in each iterat
ion), substantial progress has been made recently towards understanding the samp
le efficiency of Q-learning. To yield an entrywise $\varepsilon$-accurate estima
te of the optimal Q-function, state-of-the-art theory requires at least an order
 of $\frac{|S||A|}{(1-\gamma)^5\varepsilon^{2}}$ samples in the infinite-horizon
 $\gamma$-discounted setting. In this work, we sharpen the sample complexity of
synchronous Q-learning to the order of $\frac{|S||A|}{(1-\gamma)^4\varepsilon^2}
$ (up to some logarithmic factor) for any $0
```
Copy to ClipboardDownload

Endnote
%0 Conference Paper
%T Tightening the Dependence on Horizon in the Sample Complexity of Q-Learning
%A Gen Li
%A Changxiao Cai
%A Yuxin Chen
%A Yuantao Gu
%A Yuting Wei
%A Yuejie Chi
%B Proceedings of the 38th International Conference on Machine Learning
%C Proceedings of Machine Learning Research
%D 2021
%E Marina Meila
%E Tong Zhang■
%F pmlr-v139-li21b
%I PMLR
%P 6296--6306
%U https://proceedings.mlr.press/v139/li21b.html
%V 139
%X Q-learning, which seeks to learn the optimal Q-function of a Markov decision process (MDP) in a model-free fashion, lies at the heart of reinforcement learning. Focusing on the synchronous setting (such that independent samples for all state-action pairs are queried via a generative model in each iteration), substantial progress has been made recently towards understanding the sample efficiency of Q-learning. To yield an entrywise ε-accurate estimate of the optimal Q-function, state-of-the-art theory requires at least an order of $\frac{|S||A|}{(1-\gamma)^5\varepsilon^{2}}$ samples in the infinite-horizon γ-discounted setting. In this work, we sharpen the sample complexity of synchronous Q-learning to the order of $\frac{|S||A|}{(1-\gamma)^4\varepsilon^2}$ (up to some logarithmic factor) for any $0
Copy to ClipboardDownload
APA

Li, G., Cai, C., Chen, Y., Gu, Y., Wei, Y. & Chi, Y.. (2021). Tightening the Dependence on Horizon in the Sample Complexity of Q-Learning. Proceedings of the 38th International Conference on Machine Learning, in Proceedings of Machine Learning Research 139:6296-6306 Available from https://proceedings.mlr.press/v139/li21b.html.

Copy to ClipboardDownload

Related Material

Download PDF
Supplementary PDF
****
Winograd Algorithm for AdderNet
Wenshuo Li, Hanting Chen, Mingqiang Huang, Xinghao Chen, Chunjing Xu, Yunhe Wang
Adder neural network (AdderNet) is a new kind of deep model that replaces the original massive multiplications in convolutions by additions while preserving the high performance. Since the hardware complexity of additions is much lower than that of multiplications, the overall energy consumption is thus reduced significantly. To further optimize the hardware overhead of using AdderNet, this paper studies the winograd algorithm, which is a widely used fast algorithm for accelerating convolution and saving the computational costs. Unfortunately, the conventional Winograd algorithm cannot be directly applied to AdderNets since the distributive law in multiplication is not valid for the l1-norm. Therefore, we replace the element-wise multiplication in the Winograd equation by additions and the

n develop a new set of transform matrixes that can enhance the representation ability of output features to maintain the performance. Moreover, we propose the l2-to-l1 training strategy to mitigate the negative impacts caused by formal inconsistency. Experimental results on both FPGA and benchmarks show that the new method can further reduce the energy consumption without affecting the accuracy of the original AdderNet.
****
A Free Lunch From ANN: Towards Efficient, Accurate Spiking Neural Networks Calibration
Yuhang Li, Shikuang Deng, Xin Dong, Ruihao Gong, Shi Gu
Spiking Neural Network (SNN) has been recognized as one of the next generation of neural networks. Conventionally, SNN can be converted from a pre-trained ANN by only replacing the ReLU activation to spike activation while keeping the parameters intact. Perhaps surprisingly, in this work we show that a proper way to calibrate the parameters during the conversion of ANN to SNN can bring significant improvements. We introduce SNN Calibration, a cheap but extraordinarily effective method by leveraging the knowledge within a pre-trained Artificial Neural Network (ANN). Starting by analyzing the conversion error and its propagation through layers theoretically, we propose the calibration algorithm that can correct the error layer-by-layer. The calibration only takes a handful number of training data and several minutes to finish. Moreover, our calibration algorithm can produce SNN with state-of-the-art architecture on the large-scale ImageNet dataset, including MobileNet and RegNet. Extensive experiments demonstrate the effectiveness and efficiency of our algorithm. For example, our advanced pipeline can increase up to 69% top-1 accuracy when converting MobileNet on ImageNet compared to baselines. Codes are released at https://github.com/yhhhli/SNN_Calibration.
****
Privacy-Preserving Feature Selection with Secure Multiparty Computation
Xiling Li, Rafael Dowsley, Martine De Cock
Existing work on privacy-preserving machine learning with Secure Multiparty Computation (MPC) is almost exclusively focused on model training and on inference with trained models, thereby overlooking the important data pre-processing stage. In this work, we propose the first MPC based protocol for private feature selection based on the filter method, which is independent of model training, and can be used in combination with any MPC protocol to rank features. We propose an efficient feature scoring protocol based on Gini impurity to this end. To demonstrate the feasibility of our approach for practical data science, we perform experiments with the proposed MPC protocols for feature selection in a commonly used machine-learning-as-a-service configuration where computations are outsourced to multiple servers, with semi-honest and with malicious adversaries. Regarding effectiveness, we show that secure feature selection with the proposed protocols improves the accuracy of classifiers on a variety of real-world data sets, without leaking information about the feature values or even which features were selected. Regarding efficiency, we document runtimes ranging from several seconds to an hour for our protocols to finish, depending on the size of the data set and the security settings.
****
Theory of Spectral Method for Union of Subspaces-Based Random Geometry Graph
Gen Li, Yuantao Gu
Spectral method is a commonly used scheme to cluster data points lying close to Union of Subspaces, a task known as Subspace Clustering. The typical usage is to construct a Random Geometry Graph first and then apply spectral method to the graph to obtain clustering result. The latter step has been coined the name Spectral Clustering. As far as we know, in spite of the significance of both steps in spectral-method-based Subspace Clustering, all existing theoretical results focus on the first step of constructing the graph, but ignore the final step to correct false connections through spectral clustering. This paper establishes a theory to show the power of this method for the first time, in which we demonstrate the mechanism of spectral clustering by analyzing a simplified algorithm under the widely used semi-random model. Based on this theory, we prove the efficiency

of Subspace Clustering in fairly broad conditions. The insights and analysis techniques developed in this paper might also have implications for other random graph problems.
****
MURAL: Meta-Learning Uncertainty-Aware Rewards for Outcome-Driven Reinforcement Learning
Kevin Li, Abhishek Gupta, Ashwin Reddy, Vitchyr H Pong, Aurick Zhou, Justin Yu, Sergey Levine
Exploration in reinforcement learning is, in general, a challenging problem. A common technique to make learning easier is providing demonstrations from a human supervisor, but such demonstrations can be expensive and time-consuming to acquire. In this work, we study a more tractable class of reinforcement learning problems defined simply by examples of successful outcome states, which can be much easier to provide while still making the exploration problem more tractable. In this problem setting, the reward function can be obtained automatically by training a classifier to categorize states as successful or not. However, as we will show, this requires the classifier to make uncertainty-aware predictions that are very difficult using standard techniques for training deep networks. To address this, we propose a novel mechanism for obtaining calibrated uncertainty based on an amortized technique for computing the normalized maximum likelihood (NML) distribution, leveraging tools from meta-learning to make this distribution tractable. We show that the resulting algorithm has a number of intriguing connections to both count-based exploration methods and prior algorithms for learning reward functions, while also providing more effective guidance towards the goal. We demonstrate that our algorithm solves a number of challenging navigation and robotic manipulation tasks which prove difficult or impossible for prior methods.
****
Ditto: Fair and Robust Federated Learning Through Personalization
Tian Li, Shengyuan Hu, Ahmad Beirami, Virginia Smith
Fairness and robustness are two important concerns for federated learning systems. In this work, we identify that robustness to data and model poisoning attacks and fairness, measured as the uniformity of performance across devices, are competing constraints in statistically heterogeneous networks. To address these constraints, we propose employing a simple, general framework for personalized federated learning, Ditto, that can inherently provide fairness and robustness benefits, and develop a scalable solver for it. Theoretically, we analyze the ability of Ditto to achieve fairness and robustness simultaneously on a class of linear problems. Empirically, across a suite of federated datasets, we show that Ditto not only achieves competitive performance relative to recent personalization methods, but also enables more accurate, robust, and fair models relative to state-of-the-art fair or robust baselines.
****
Quantization Algorithms for Random Fourier Features
Xiaoyun Li, Ping Li
The method of random projection (RP) is the standard technique for dimensionality reduction, approximate near neighbor search, compressed sensing, etc., which provides a simple and effective scheme for approximating pairwise inner products and Euclidean distances in massive data. Closely related to RP, the method of random Fourier features (RFF) has also become popular for approximating the (nonlinear) Gaussian kernel. RFF applies a specific nonlinear transformation on the projected data from RP. In practice, using the Gaussian kernel often leads to better performance than the linear kernel (inner product). After random projections, quantization is an important step for efficient data storage, computation and transmission. Quantization for RP has been extensively studied in the literature. In this paper, we focus on developing quantization algorithms for RFF. The task is in a sense challenging due to the tuning parameter γ in the Gaussian kernel. For example, the quantizer and the quantized data might be tied to each specific Gaussian kernel parameter γ. Our contribution begins with the analysis on the probability distributions of RFF, and an interesting discovery that the marginal distribution of RFF is free of the parameter γ. This signi

ficantly simplifies the design of the Lloyd-Max (LM) quantization scheme for RFF in that there would be only one LM quantizer (regardless of γ). Detailed theoretical analysis is provided on the kernel estimators and approximation error, and experiments confirm the effectiveness and efficiency of the proposed method.
****
Approximate Group Fairness for Clustering
Bo Li, Lijun Li, Ankang Sun, Chenhao Wang, Yingfan Wang
We incorporate group fairness into the algorithmic centroid clustering problem, where k centers are to be located to serve n agents distributed in a metric space. We refine the notion of proportional fairness proposed in [Chen et al., ICML 2019] as {\em core fairness}. A k-clustering is in the core if no coalition containing at least n/k agents can strictly decrease their total distance by deviating to a new center together. Our solution concept is motivated by the situation where agents are able to coordinate and utilities are transferable. A string of existence, hardness and approximability results is provided. Particularly, we propose two dimensions to relax core requirements: one is on the degree of distance improvement, and the other is on the size of deviating coalition. For both relaxations and their combination, we study the extent to which relaxed core fairness can be satisfied in metric spaces including line, tree and general metric space, and design approximation algorithms accordingly. We also conduct experiments on synthetic and real-world data to examine the performance of our algorithms.
****
Sharper Generalization Bounds for Clustering
Shaojie Li, Yong Liu
Existing generalization analysis of clustering mainly focuses on specific instantiations, such as (kernel) k-means, and a unified framework for studying clustering performance is still lacking. Besides, the existing excess clustering risk bounds are mostly of order $\mathcal{O}(K/\sqrt{n})$ provided that the underlying distribution has bounded support, where n is the sample size and K is the cluster numbers, or of order $\mathcal{O}(K^2/n)$ under strong assumptions on the underlying distribution, where these assumptions are hard to be verified in general. In this paper, we propose a unified clustering learning framework and investigate its excess risk bounds, obtaining state-of-the-art upper bounds under mild assumptions. Specifically, we derive sharper bounds of order $\mathcal{O}(K^2/n)$ under mild assumptions on the covering number of the hypothesis spaces, where these assumptions are easy to be verified. Moreover, for the hard clustering scheme, such as (kernel) k-means, if just assume the hypothesis functions to be bounded, we improve the upper bounds from the order $\mathcal{O}(K/\sqrt{n})$ to $\mathcal{O}(\sqrt{K}/\sqrt{n})$. Furthermore, state-of-the-art bounds of faster order $\mathcal{O}(K/n)$ are obtained with the covering number assumptions.
****
Provably End-to-end Label-noise Learning without Anchor Points
Xuefeng Li, Tongliang Liu, Bo Han, Gang Niu, Masashi Sugiyama
In label-noise learning, the transition matrix plays a key role in building statistically consistent classifiers. Existing consistent estimators for the transition matrix have been developed by exploiting anchor points. However, the anchor-point assumption is not always satisfied in real scenarios. In this paper, we propose an end-to-end framework for solving label-noise learning without anchor points, in which we simultaneously optimize two objectives: the cross entropy loss between the noisy label and the predicted probability by the neural network, and the volume of the simplex formed by the columns of the transition matrix. Our proposed framework can identify the transition matrix if the clean class-posterior probabilities are sufficiently scattered. This is by far the mildest assumption under which the transition matrix is provably identifiable and the learned classifier is statistically consistent. Experimental results on benchmark datasets demonstrate the effectiveness and robustness of the proposed method.
****

A Novel Method to Solve Neural Knapsack Problems

Duanshun Li, Jing Liu, Dongeun Lee, Ali Seyedmazloom, Giridhar Kaushik, Kookjin Lee, Noseong Park

0-1 knapsack is of fundamental importance across many fields. In this paper, we present a game-theoretic method to solve 0-1 knapsack problems (KPs) where the number of items (products) is large and the values of items are not predetermined but decided by an external value assignment function (e.g., a neural network in our case) during the optimization process. While existing papers are interested in predicting solutions with neural networks for classical KPs whose objective functions are mostly linear functions, we are interested in solving KPs whose objective functions are neural networks. In other words, we choose a subset of items that maximize the sum of the values predicted by neural networks. Its key challenge is how to optimize the neural network-based non-linear KP objective with a budget constraint. Our solution is inspired by game-theoretic approaches in deep learning, e.g., generative adversarial networks. After formally defining our two-player game, we develop an adaptive gradient ascent method to solve it. In our experiments, our method successfully solves two neural network-based non-linear KPs and conventional linear KPs with 1 million items.

****

Mixed Cross Entropy Loss for Neural Machine Translation

Haoran Li, Wei Lu

In neural machine translation, Cross Entropy loss (CE) is the standard loss function in two training methods of auto-regressive models, i.e., teacher forcing and scheduled sampling. In this paper, we propose mixed Cross Entropy loss (mixed CE) as a substitute for CE in both training approaches. In teacher forcing, the model trained with CE regards the translation problem as a one-to-one mapping process, while in mixed CE this process can be relaxed to one-to-many. In scheduled sampling, we show that mixed CE has the potential to encourage the training and testing behaviours to be similar to each other, more effectively mitigating the exposure bias problem. We demonstrate the superiority of mixed CE over CE on several machine translation datasets, WMT’16 Ro-En, WMT’16 Ru-En, and WMT’14 En-De in both teacher forcing and scheduled sampling setups. Furthermore, in WMT’14 En-De, we also find mixed CE consistently outperforms CE on a multi-reference set as well as a challenging paraphrased reference set. We also found the model trained with mixed CE is able to provide a better probability distribution defined over the translation output space. Our code is available at https://github.com/haorannlp/mix.

****

Training Graph Neural Networks with 1000 Layers

Guohao Li, Matthias Müller, Bernard Ghanem, Vladlen Koltun

Deep graph neural networks (GNNs) have achieved excellent results on various tasks on increasingly large graph datasets with millions of nodes and edges. However, memory complexity has become a major obstacle when training deep GNNs for practical applications due to the immense number of nodes, edges, and intermediate activations. To improve the scalability of GNNs, prior works propose smart graph sampling or partitioning strategies to train GNNs with a smaller set of nodes or sub-graphs. In this work, we study reversible connections, group convolutions, weight tying, and equilibrium models to advance the memory and parameter efficiency of GNNs. We find that reversible connections in combination with deep network architectures enable the training of overparameterized GNNs that significantly outperform existing methods on multiple datasets. Our models RevGNN-Deep (1001 layers with 80 channels each) and RevGNN-Wide (448 layers with 224 channels each) were both trained on a single commodity GPU and achieve an ROC-AUC of 87.74 $\pm$ 0.13 and 88.14 $\pm$ 0.15 on the ogbn-proteins dataset. To the best of our knowledge, RevGNN-Deep is the deepest GNN in the literature by one order of magnitude.

****

Active Feature Acquisition with Generative Surrogate Models

Yang Li, Junier Oliva

Many real-world situations allow for the acquisition of additional relevant info

rmation when making an assessment with limited or uncertain data. However, tradi
tional ML approaches either require all features to be acquired beforehand or re
gard part of them as missing data that cannot be acquired. In this work, we cons
ider models that perform active feature acquisition (AFA) and query the environm
ent for unobserved features to improve the prediction assessments at evaluation
time. Our work reformulates the Markov decision process (MDP) that underlies the
 AFA problem as a generative modeling task and optimizes a policy via a novel mo
del-based approach. We propose learning a generative surrogate model (GSM) that
captures the dependencies among input features to assess potential information g
ain from acquisitions. The GSM is leveraged to provide intermediate rewards and
auxiliary information to aid the agent navigate a complicated high-dimensional a
ction space and sparse rewards. Furthermore, we extend AFA in a task we coin act
ive instance recognition (AIR) for the unsupervised case where the target variab
les are the unobserved features themselves and the goal is to collect informatio
n for a particular instance in a cost-efficient way. Empirical results demonstra
te that our approach achieves considerably better performance than previous stat
e of the art methods on both supervised and unsupervised tasks.
****
Partially Observed Exchangeable Modeling
Yang Li, Junier Oliva
Modeling dependencies among features is fundamental for many machine learning ta
sks. Although there are often multiple related instances that may be leveraged t
o inform conditional dependencies, typical approaches only model conditional dep
endencies over individual instances. In this work, we propose a novel framework,
 partially observed exchangeable modeling (POEx) that takes in a set of related
partially observed instances and infers the conditional distribution for the uno
bserved dimensions over multiple elements. Our approach jointly models the intra
-instance (among features in a point) and inter-instance (among multiple points
in a set) dependencies in data. POEx is a general framework that encompasses man
y existing tasks such as point cloud expansion and few-shot generation, as well
as new tasks like few-shot imputation. Despite its generality, extensive empiric
al evaluations show that our model achieves state-of-the-art performance across
a range of applications.
****
Testing DNN-based Autonomous Driving Systems under Critical Environmental Condit
ions
Zhong Li, Minxue Pan, Tian Zhang, Xuandong Li
Due to the increasing usage of Deep Neural Network (DNN) based autonomous drivin
g systems (ADS) where erroneous or unexpected behaviours can lead to catastrophi
c accidents, testing such systems is of growing importance. Existing approaches
often just focus on finding erroneous behaviours and have not thoroughly studied
 the impact of environmental conditions. In this paper, we propose to test DNN-b
ased ADS under different environmental conditions to identify the critical ones,
 that is, the environmental conditions under which the ADS are more prone to err
ors. To tackle the problem of the space of environmental conditions being extrem
ely large, we present a novel approach named TACTIC that employs the search-base
d method to identify critical environmental conditions generated by an image-to-
image translation model. Large-scale experiments show that TACTIC can effectivel
y identify critical environmental conditions and produce realistic testing image
s, and meanwhile, reveal more erroneous behaviours compared to existing approach
es.
****
The Symmetry between Arms and Knapsacks: A Primal-Dual Approach for Bandits with
 Knapsacks
Xiaocheng Li, Chunlin Sun, Yinyu Ye
In this paper, we study the bandits with knapsacks (BwK) problem and develop a p
rimal-dual based algorithm that achieves a problem-dependent logarithmic regret
bound. The BwK problem extends the multi-arm bandit (MAB) problem to model the r
esource consumption, and the existing BwK literature has been mainly focused on
deriving asymptotically optimal distribution-free regret bounds. We first study

the primal and dual linear programs underlying the BwK problem. From this primal -dual perspective, we discover symmetry between arms and knapsacks, and then pro pose a new notion of suboptimality measure for the BwK problem. The suboptimalit y measure highlights the important role of knapsacks in determining algorithm re gret and inspires the design of our two-phase algorithm. In the first phase, the algorithm identifies the optimal arms and the binding knapsacks, and in the sec ond phase, it exhausts the binding knapsacks via playing the optimal arms throug h an adaptive procedure. Our regret upper bound involves the proposed suboptimal ity measure and it has a logarithmic dependence on length of horizon T and a p olynomial dependence on m (the numbers of arms) and d (the number of knapsac ks). To the best of our knowledge, this is the first problem-dependent logarithm ic regret bound for solving the general BwK problem.
****
Distributionally Robust Optimization with Markovian Data
Mengmeng Li, Tobias Sutter, Daniel Kuhn
We study a stochastic program where the probability distribution of the uncertai n problem parameters is unknown and only indirectly observed via finitely many c orrelated samples generated by an unknown Markov chain with d states. We propo se a data-driven distributionally robust optimization model to estimate the prob lem's objective function and optimal solution. By leveraging results from large deviations theory, we derive statistical guarantees on the quality of these esti mators. The underlying worst-case expectation problem is nonconvex and involves $\mathcal O(d^2)$ decision variables. Thus, it cannot be solved efficiently for large d. By exploiting the structure of this problem, we devise a customized F rank-Wolfe algorithm with convex direction-finding subproblems of size $\mathcal O(d)$. We prove that this algorithm finds a stationary point efficiently under mild conditions. The efficiency of the method is predicated on a dimensionality reduction enabled by a dual reformulation. Numerical experiments indicate that o ur approach has better computational and statistical properties than the state-o f-the-art methods.
****
Communication-Efficient Distributed SVD via Local Power Iterations
Xiang Li, Shusen Wang, Kun Chen, Zhihua Zhang
We study distributed computing of the truncated singular value decomposition (SV D). We develop an algorithm that we call \texttt{LocalPower} for improving commu nication efficiency. Specifically, we uniformly partition the dataset among m nodes and alternate between multiple (precisely p) local power iterations and one global aggregation. In the aggregation, we propose to weight each local eige nvector matrix with orthogonal Procrustes transformation (OPT). As a practical s urrogate of OPT, sign-fixing, which uses a diagonal matrix with ± 1 entries as weights, has better computation complexity and stability in experiments. We t heoretically show that under certain assumptions \texttt{LocalPower} lowers the required number of communications by a factor of p to reach a constant accurac y. We also show that the strategy of periodically decaying p helps obtain high -precision solutions. We conduct experiments to demonstrate the effectiveness of \texttt{LocalPower}.
****
FILTRA: Rethinking Steerable CNN by Filter Transform
Bo Li, Qili Wang, Gim Hee Lee
Steerable CNN imposes the prior knowledge of transformation invariance or equiva riance in the network architecture to enhance the the network robustness on geom etry transformation of data and reduce overfitting. It has been an intuitive and widely used technique to construct a steerable filter by augmenting a filter wi th its transformed copies in the past decades, which is named as filter transfor m in this paper. Recently, the problem of steerable CNN has been studied from as pect of group representation theory, which reveals the function space structure of a steerable kernel function. However, it is not yet clear on how this theory is related to the filter transform technique. In this paper, we show that kernel constructed by filter transform can also be interpreted in the group representa tion theory. This interpretation help complete the puzzle of steerable CNN theor

y and provides a novel and simple approach to implement steerable convolution operators. Experiments are executed on multiple datasets to verify the feasibility of the proposed approach.
****
Online Unrelated Machine Load Balancing with Predictions Revisited
Shi Li, Jiayi Xian
We study the online load balancing problem with machine learned predictions, and give results that improve upon and extend those in a recent paper by Lattanzi et al. (2020). First, we design deterministic and randomized online rounding algorithms for the problem in the unrelated machine setting, with $O(\frac{\log m}{\log \log m})$- and $O(\frac{\log \log m}{\log \log \log m})$-competitive ratios. They respectively improve upon the previous ratios of $O(\log m)$ and $O(\log^3 \log m)$, and match the lower bounds given by Lattanzi et al. Second, we extend their prediction scheme from the identical machine restricted assignment setting to the unrelated machine setting. With the knowledge of two vectors over machines, a dual vector and a weight vector, we can construct a good fractional assignment online, that can be passed to an online rounding algorithm. Finally, we consider the learning model introduced by Lavastida et al. (2020), and show that under the model, the two vectors can be learned efficiently with a few samples of instances.
****
Asymptotic Normality and Confidence Intervals for Prediction Risk of the Min-Norm Least Squares Estimator
Zeng Li, Chuanlong Xie, Qinwen Wang
This paper quantifies the uncertainty of prediction risk for the min-norm least squares estimator in high-dimensional linear regression models. We establish the asymptotic normality of prediction risk when both the sample size and the number of features tend to infinity. Based on the newly established central limit theorems(CLTs), we derive the confidence intervals of the prediction risk under various scenarios. Our results demonstrate the sample-wise non-monotonicity of the prediction risk and confirm “more data hurt" phenomenon. Furthermore, the width of confidence intervals indicates that over-parameterization would enlarge the randomness of prediction performance.
****
TeraPipe: Token-Level Pipeline Parallelism for Training Large-Scale Language Models
Zhuohan Li, Siyuan Zhuang, Shiyuan Guo, Danyang Zhuo, Hao Zhang, Dawn Song, Ion Stoica
Model parallelism has become a necessity for training modern large-scale deep language models. In this work, we identify a new and orthogonal dimension from existing model parallel approaches: it is possible to perform pipeline parallelism within a single training sequence for Transformer-based language models thanks to its autoregressive property. This enables a more fine-grained pipeline compared with previous work. With this key idea, we design TeraPipe, a high-performance token-level pipeline parallel algorithm for synchronous model-parallel training of Transformer-based language models. We develop a novel dynamic programming-based algorithm to calculate the optimal pipelining execution scheme given a specific model and cluster configuration. We show that TeraPipe can speed up the training by 5.0x for the largest GPT-3 model with 175 billion parameters on an AWS cluster with 48 p3.16xlarge instances compared with state-of-the-art model-parallel methods. The code for reproduction can be found at https://github.com/zhuohan123/terapipe
****
A Second look at Exponential and Cosine Step Sizes: Simplicity, Adaptivity, and Performance
Xiaoyu Li, Zhenxun Zhuang, Francesco Orabona
Stochastic Gradient Descent (SGD) is a popular tool in training large-scale machine learning models. Its performance, however, is highly variable, depending crucially on the choice of the step sizes. Accordingly, a variety of strategies for tuning the step sizes have been proposed, ranging from coordinate-wise approach

es (a.k.a. “adaptive” step sizes) to sophisticated heuristics to change the step size in each iteration. In this paper, we study two step size schedules whose power has been repeatedly confirmed in practice: the exponential and the cosine step sizes. For the first time, we provide theoretical support for them proving convergence rates for smooth non-convex functions, with and without the Polyak-■{}ojasiewicz (PL) condition. Moreover, we show the surprising property that these two strategies are \emph{adaptive} to the noise level in the stochastic gradients of PL functions. That is, contrary to polynomial step sizes, they achieve almost optimal performance without needing to know the noise level nor tuning their hyperparameters based on it. Finally, we conduct a fair and comprehensive empirical evaluation of real-world datasets with deep learning architectures. Results show that, even if only requiring at most two hyperparameters to tune, these two strategies best or match the performance of various finely-tuned state-of-the-art strategies.
****
Towards Understanding and Mitigating Social Biases in Language Models
Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, Ruslan Salakhutdinov
As machine learning methods are deployed in real-world settings such as healthcare, legal systems, and social science, it is crucial to recognize how they shape social biases and stereotypes in these sensitive decision-making processes. Among such real-world deployments are large-scale pretrained language models (LMs) that can be potentially dangerous in manifesting undesirable representational biases - harmful biases resulting from stereotyping that propagate negative generalizations involving gender, race, religion, and other social constructs. As a step towards improving the fairness of LMs, we carefully define several sources of representational biases before proposing new benchmarks and metrics to measure them. With these tools, we propose steps towards mitigating social biases during text generation. Our empirical results and human evaluation demonstrate effectiveness in mitigating bias while retaining crucial contextual information for high-fidelity text generation, thereby pushing forward the performance-fairness Pareto frontier.
****
Uncovering the Connections Between Adversarial Transferability and Knowledge Transferability
Kaizhao Liang, Jacky Y Zhang, Boxin Wang, Zhuolin Yang, Sanmi Koyejo, Bo Li
Knowledge transferability, or transfer learning, has been widely adopted to allow a pre-trained model in the source domain to be effectively adapted to downstream tasks in the target domain. It is thus important to explore and understand the factors affecting knowledge transferability. In this paper, as the first work, we analyze and demonstrate the connections between knowledge transferability and another important phenomenon–adversarial transferability, \emph{i.e.}, adversarial examples generated against one model can be transferred to attack other models. Our theoretical studies show that adversarial transferability indicates knowledge transferability, and vice versa. Moreover, based on the theoretical insights, we propose two practical adversarial transferability metrics to characterize this process, serving as bidirectional indicators between adversarial and knowledge transferability. We conduct extensive experiments for different scenarios on diverse datasets, showing a positive correlation between adversarial transferability and knowledge transferability. Our findings will shed light on future research about effective knowledge transfer learning and adversarial transferability analyses.
****
Parallel Droplet Control in MEDA Biochips using Multi-Agent Reinforcement Learning
Tung-Che Liang, Jin Zhou, Yun-Sheng Chan, Tsung-Yi Ho, Krishnendu Chakrabarty, Cy Lee
Microfluidic biochips are being utilized for clinical diagnostics, including COVID-19 testing, because of they provide sample-to-result turnaround at low cost. Recently, microelectrode-dot-array (MEDA) biochips have been proposed to advance microfluidics technology. A MEDA biochip manipulates droplets of nano/picoliter

volumes to automatically execute biochemical protocols. During bioassay execution, droplets are transported in parallel to achieve high-throughput outcomes. However, a major concern associated with the use of MEDA biochips is microelectrode degradation over time. Recent work has shown that formulating droplet transportation as a reinforcement-learning (RL) problem enables the training of policies to capture the underlying health conditions of microelectrodes and ensure reliable fluidic operations. However, the above RL-based approach suffers from two key limitations: 1) it cannot be used for concurrent transportation of multiple droplets; 2) it requires the availability of CCD cameras for monitoring droplet movement. To overcome these problems, we present a multi-agent reinforcement learning (MARL) droplet-routing solution that can be used for various sizes of MEDA biochips with integrated sensors, and we demonstrate the reliable execution of a serial-dilution bioassay with the MARL droplet router on a fabricated MEDA biochip. To facilitate further research, we also present a simulation environment based on the PettingZoo Gym Interface for MARL-guided droplet-routing problems on MEDA biochips.
****
Information Obfuscation of Graph Neural Networks
Peiyuan Liao, Han Zhao, Keyulu Xu, Tommi Jaakkola, Geoffrey J. Gordon, Stefanie Jegelka, Ruslan Salakhutdinov
While the advent of Graph Neural Networks (GNNs) has greatly improved node and graph representation learning in many applications, the neighborhood aggregation scheme exposes additional vulnerabilities to adversaries seeking to extract node-level information about sensitive attributes. In this paper, we study the problem of protecting sensitive attributes by information obfuscation when learning with graph structured data. We propose a framework to locally filter out pre-determined sensitive attributes via adversarial training with the total variation and the Wasserstein distance. Our method creates a strong defense against inference attacks, while only suffering small loss in task performance. Theoretically, we analyze the effectiveness of our framework against a worst-case adversary, and characterize an inherent trade-off between maximizing predictive accuracy and minimizing information leakage. Experiments across multiple datasets from recommender systems, knowledge graphs and quantum chemistry demonstrate that the proposed approach provides a robust defense across various graph structures and tasks, while producing competitive GNN encoders for downstream tasks.
****
Guided Exploration with Proximal Policy Optimization using a Single Demonstration
Gabriele Libardi, Gianni De Fabritiis, Sebastian Dittert
Solving sparse reward tasks through exploration is one of the major challenges in deep reinforcement learning, especially in three-dimensional, partially-observable environments. Critically, the algorithm proposed in this article is capable of using a single human demonstration to solve hard-exploration problems. We train an agent on a combination of demonstrations and own experience to solve problems with variable initial conditions and we integrate it with proximal policy optimization (PPO). The agent is also able to increase its performance and to tackle harder problems by replaying its own past trajectories prioritizing them based on the obtained reward and the maximum value of the trajectory. We finally compare variations of this algorithm to different imitation learning algorithms on a set of hard-exploration tasks in the Animal-AI Olympics environment. To the best of our knowledge, learning a task in a three-dimensional environment with comparable difficulty has never been considered before using only one human demonstration.
****
Debiasing a First-order Heuristic for Approximate Bi-level Optimization
Valerii Likhosherstov, Xingyou Song, Krzysztof Choromanski, Jared Q Davis, Adrian Weller
Approximate bi-level optimization (ABLO) consists of (outer-level) optimization problems, involving numerical (inner-level) optimization loops. While ABLO has many applications across deep learning, it suffers from time and memory complexit

y proportional to the length r of its inner optimization loop. To address this complexity, an earlier first-order method (FOM) was proposed as a heuristic which omits second derivative terms, yielding significant speed gains and requiring only constant memory. Despite FOM's popularity, there is a lack of theoretical understanding of its convergence properties. We contribute by theoretically characterizing FOM's gradient bias under mild assumptions. We further demonstrate a rich family of examples where FOM-based SGD does not converge to a stationary point of the ABLO objective. We address this concern by proposing an unbiased FOM (UFOM) enjoying constant memory complexity as a function of r. We characterize the introduced time-variance tradeoff, demonstrate convergence bounds, and find an optimal UFOM for a given ABLO problem. Finally, we propose an efficient adaptive UFOM scheme.
****
Making transport more robust and interpretable by moving data through a small number of anchor points
Chi-Heng Lin, Mehdi Azabou, Eva Dyer
Optimal transport (OT) is a widely used technique for distribution alignment, with applications throughout the machine learning, graphics, and vision communities. Without any additional structural assumptions on transport, however, OT can be fragile to outliers or noise, especially in high dimensions. Here, we introduce Latent Optimal Transport (LOT), a new approach for OT that simultaneously learns low-dimensional structure in data while leveraging this structure to solve the alignment task. The idea behind our approach is to learn two sets of "anchors" that constrain the flow of transport between a source and target distribution. In both theoretical and empirical studies, we show that LOT regularizes the rank of transport and makes it more robust to outliers and the sampling density. We show that by allowing the source and target to have different anchors, and using LOT to align the latent spaces between anchors, the resulting transport plan has better structural interpretability and highlights connections between both the individual data points and the local geometry of the datasets.
****
Straight to the Gradient: Learning to Use Novel Tokens for Neural Text Generation
Xiang Lin, Simeng Han, Shafiq Joty
Advanced large-scale neural language models have led to significant success in many language generation tasks. However, the most commonly used training objective, Maximum Likelihood Estimation (MLE), has been shown problematic, where the trained model prefers using dull and repetitive phrases. In this work, we introduce ScaleGrad, a modification straight to the gradient of the loss function, to remedy the degeneration issue of the standard MLE objective. By directly maneuvering the gradient information, ScaleGrad makes the model learn to use novel tokens. Empirical results show the effectiveness of our method not only in open-ended generation, but also in directed generation tasks. With the simplicity in architecture, our method can serve as a general training objective that is applicable to most of the neural text generation tasks.
****
Quasi-global Momentum: Accelerating Decentralized Deep Learning on Heterogeneous Data
Tao Lin, Sai Praneeth Karimireddy, Sebastian Stich, Martin Jaggi
Decentralized training of deep learning models is a key element for enabling data privacy and on-device learning over networks. In realistic learning scenarios, the presence of heterogeneity across different clients' local datasets poses an optimization challenge and may severely deteriorate the generalization performance. In this paper, we investigate and identify the limitation of several decentralized optimization algorithms for different degrees of data heterogeneity. We propose a novel momentum-based method to mitigate this decentralized training difficulty. We show in extensive empirical experiments on various CV/NLP datasets (CIFAR-10, ImageNet, and AG News) and several network topologies (Ring and Social Network) that our method is much more robust to the heterogeneity of clients' data than other existing methods, by a significant improvement in test performan

ce (1%-20%).
****
Generative Causal Explanations for Graph Neural Networks
Wanyu Lin, Hao Lan, Baochun Li
This paper presents {\em Gem}, a model-agnostic approach for providing interpretable explanations for any GNNs on various graph learning tasks. Specifically, we formulate the problem of providing explanations for the decisions of GNNs as a causal learning task. Then we train a causal explanation model equipped with a loss function based on Granger causality. Different from existing explainers for GNNs, {\em Gem} explains GNNs on graph-structured data from a causal perspective. It has better generalization ability as it has no requirements on the internal structure of the GNNs or prior knowledge on the graph learning tasks. In addition, {\em Gem}, once trained, can be used to explain the target GNN very quickly. Our theoretical analysis shows that several recent explainers fall into a unified framework of {\em additive feature attribution methods}. Experimental results on synthetic and real-world datasets show that {\em Gem} achieves a relative increase of the explanation accuracy by up to $30%$ and speeds up the explanation process by up to $110\times$ as compared to its state-of-the-art alternatives.
****
Tractable structured natural-gradient descent using local parameterizations
Wu Lin, Frank Nielsen, Khan Mohammad Emtiyaz, Mark Schmidt
Natural-gradient descent (NGD) on structured parameter spaces (e.g., low-rank covariances) is computationally challenging due to difficult Fisher-matrix computations. We address this issue by using \emph{local-parameter coordinates} to obtain a flexible and efficient NGD method that works well for a wide-variety of structured parameterizations. We show four applications where our method (1) generalizes the exponential natural evolutionary strategy, (2) recovers existing Newton-like algorithms, (3) yields new structured second-order algorithms, and (4) gives new algorithms to learn covariances of Gaussian and Wishart-based distributions. We show results on a range of problems from deep learning, variational inference, and evolution strategies. Our work opens a new direction for scalable structured geometric methods.
****
Active Learning of Continuous-time Bayesian Networks through Interventions
Dominik Linzner, Heinz Koeppl
We consider the problem of learning structures and parameters of Continuous-time Bayesian Networks (CTBNs) from time-course data under minimal experimental resources. In practice, the cost of generating experimental data poses a bottleneck, especially in the natural and social sciences. A popular approach to overcome this is Bayesian optimal experimental design (BOED). However, BOED becomes infeasible in high-dimensional settings, as it involves integration over all possible experimental outcomes. We propose a novel criterion for experimental design based on a variational approximation of the expected information gain. We show that for CTBNs, a semi-analytical expression for this criterion can be calculated for structure and parameter learning. By doing so, we can replace sampling over experimental outcomes by solving the CTBNs master-equation, for which scalable approximations exist. This alleviates the computational burden of sampling possible experimental outcomes in high-dimensions. We employ this framework to recommend interventional sequences. In this context, we extend the CTBN model to conditional CTBNs to incorporate interventions. We demonstrate the performance of our criterion on synthetic and real-world data.
****
Phase Transitions, Distance Functions, and Implicit Neural Representations
Yaron Lipman
Representing surfaces as zero level sets of neural networks recently emerged as a powerful modeling paradigm, named Implicit Neural Representations (INRs), serving numerous downstream applications in geometric deep learning and 3D vision. Training INRs previously required choosing between occupancy and distance function representation and different losses with unknown limit behavior and/or bias. In this paper we draw inspiration from the theory of phase transitions of fluids

and suggest a loss for training INRs that learns a density function that converges to a proper occupancy function, while its log transform converges to a distance function. Furthermore, we analyze the limit minimizer of this loss showing it satisfies the reconstruction constraints and has minimal surface perimeter, a desirable inductive bias for surface reconstruction. Training INRs with this new loss leads to state-of-the-art reconstructions on a standard benchmark.
****
The Earth Mover’s Pinball Loss: Quantiles for Histogram-Valued Regression
Florian List
Although ubiquitous in the sciences, histogram data have not received much attention by the Deep Learning community. Whilst regression and classification tasks for scalar and vector data are routinely solved by neural networks, a principled approach for estimating histogram labels as a function of an input vector or image is lacking in the literature. We present a dedicated method for Deep Learning-based histogram regression, which incorporates cross-bin information and yields distributions over possible histograms, expressed by τ-quantiles of the cumulative histogram in each bin. The crux of our approach is a new loss function obtained by applying the pinball loss to the cumulative histogram, which for 1D histograms reduces to the Earth Mover’s distance (EMD) in the special case of the median ($\tau = 0.5$), and generalizes it to arbitrary quantiles. We validate our method with an illustrative toy example, a football-related task, and an astrophysical computer vision problem. We show that with our loss function, the accuracy of the predicted median histograms is very similar to the standard EMD case (and higher than for per-bin loss functions such as cross-entropy), while the predictions become much more informative at almost no additional computational cost.
****
Understanding Instance-Level Label Noise: Disparate Impacts and Treatments
Yang Liu
This paper aims to provide understandings for the effect of an over-parameterized model, e.g. a deep neural network, memorizing instance-dependent noisy labels. We first quantify the harms caused by memorizing noisy instances, and show the disparate impacts of noisy labels for sample instances with different representation frequencies. We then analyze how several popular solutions for learning with noisy labels mitigate this harm at the instance level. Our analysis reveals that existing approaches lead to disparate treatments when handling noisy instances. While higher-frequency instances often enjoy a high probability of an improvement by applying these solutions, lower-frequency instances do not. Our analysis reveals new understandings for when these approaches work, and provides theoretical justifications for previously reported empirical observations. This observation requires us to rethink the distribution of label noise across instances and calls for different treatments for instances in different regimes.
****
APS: Active Pretraining with Successor Features
Hao Liu, Pieter Abbeel
We introduce a new unsupervised pretraining objective for reinforcement learning. During the unsupervised reward-free pretraining phase, the agent maximizes mutual information between tasks and states induced by the policy. Our key contribution is a novel lower bound of this intractable quantity. We show that by reinterpreting and combining variational successor features \citep{Hansen2020Fast} with nonparametric entropy maximization \citep{liu2021behavior}, the intractable mutual information can be efficiently optimized. The proposed method Active Pretraining with Successor Feature (APS) explores the environment via nonparametric entropy maximization, and the explored data can be efficiently leveraged to learn behavior by variational successor features. APS addresses the limitations of existing mutual information maximization based and entropy maximization based unsupervised RL, and combines the best of both worlds. When evaluated on the Atari 100k data-efficiency benchmark, our approach significantly outperforms previous methods combining unsupervised pretraining with task-specific finetuning.
****

Learning by Turning: Neural Architecture Aware Optimisation

Yang Liu, Jeremy Bernstein, Markus Meister, Yisong Yue

Descent methods for deep networks are notoriously capricious: they require careful tuning of step size, momentum and weight decay, and which method will work best on a new benchmark is a priori unclear. To address this problem, this paper conducts a combined study of neural architecture and optimisation, leading to a new optimiser called Nero: the neuronal rotator. Nero trains reliably without momentum or weight decay, works in situations where Adam and SGD fail, and requires little to no learning rate tuning. Also, Nero’s memory footprint is   square root that of Adam or LAMB. Nero combines two ideas: (1) projected gradient descent over the space of balanced networks; (2) neuron-specific updates, where the step size sets the angle through which each neuron’s hyperplane turns. The paper concludes by discussing how this geometric connection between architecture and optimisation may impact theories of generalisation in deep learning.

****

Dynamic Game Theoretic Neural Optimizer

Guan-Horng Liu, Tianrong Chen, Evangelos Theodorou

The connection between training deep neural networks (DNNs) and optimal control theory (OCT) has attracted considerable attention as a principled tool of algorithmic design. Despite few attempts being made, they have been limited to architectures where the layer propagation resembles a Markovian dynamical system. This casts doubts on their flexibility to modern networks that heavily rely on non-Markovian dependencies between layers (e.g. skip connections in residual networks). In this work, we propose a novel dynamic game perspective by viewing each layer as a player in a dynamic game characterized by the DNN itself. Through this lens, different classes of optimizers can be seen as matching different types of Nash equilibria, depending on the implicit information structure of each (p)layer. The resulting method, called Dynamic Game Theoretic Neural Optimizer (DGNOpt), not only generalizes OCT-inspired optimizers to richer network class; it also motivates a new training principle by solving a multi-player cooperative game. DGNOpt shows convergence improvements over existing methods on image classification datasets with residual and inception networks. Our work marries strengths from both OCT and game theory, paving ways to new algorithmic opportunities from robust optimal control and bandit-based optimization.

****

Besov Function Approximation and Binary Classification on Low-Dimensional Manifolds Using Convolutional Residual Networks

Hao Liu, Minshuo Chen, Tuo Zhao, Wenjing Liao

Most of existing statistical theories on deep neural networks have sample complexities cursed by the data dimension and therefore cannot well explain the empirical success of deep learning on high-dimensional data. To bridge this gap, we propose to exploit the low-dimensional structures of the real world datasets and establish theoretical guarantees of convolutional residual networks (ConvResNet) in terms of function approximation and statistical recovery for binary classification problem. Specifically, given the data lying on a d-dimensional manifold isometrically embedded in $\mathbb{R}^D$, we prove that if the network architecture is properly chosen, ConvResNets can (1) approximate {\it Besov functions} on manifolds with arbitrary accuracy, and (2) learn a classifier by minimizing the empirical logistic risk, which gives an {\it excess risk} in the order of $n^{-\frac{s}{2s+2(s\vee d)}}$, where s is a smoothness parameter. This implies that the sample complexity depends on the intrinsic dimension d, instead of the data dimension D. Our results demonstrate that ConvResNets are adaptive to low-dimensional structures of data sets.

****

Just Train Twice: Improving Group Robustness without Training Group Information

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, Chelsea Finn

Standard training via empirical risk minimization (ERM) can produce models that achieve low error on average but high error on minority groups, especially in the presence of spurious correlations between the input and label. Prior approache

s to this problem, like group distributionally robust optimization (group DRO), generally require group annotations for every training point. On the other hand, approaches that do not use group annotations generally do not improve minority performance. For example, we find that joint DRO, which dynamically upweights examples with high training loss, tends to optimize for examples that are irrelevant to the specific groups we seek to do well on. In this paper, we propose a simple two-stage approach, JTT, that achieves comparable performance to group DRO while only requiring group annotations on a significantly smaller validation set. JTT first attempts to identify informative training examples, which are often minority examples, by training an initial ERM classifier and selecting the examples with high training loss. Then, it trains a final classifier by upsampling the selected examples. Crucially, unlike joint DRO, JTT does not iteratively upsample examples that have high loss under the final classifier. On four image classification and natural language processing tasks with spurious correlations, we show that JTT closes 85% of the gap in accuracy on the worst group between ERM and group DRO.
****
Event Outlier Detection in Continuous Time
Siqi Liu, Milos Hauskrecht
Continuous-time event sequences represent discrete events occurring in continuous time. Such sequences arise frequently in real-life. Usually we expect the sequences to follow some regular pattern over time. However, sometimes these patterns may be interrupted by unexpected absence or occurrences of events. Identification of these unexpected cases can be very important as they may point to abnormal situations that need human attention. In this work, we study and develop methods for detecting outliers in continuous-time event sequences, including unexpected absence and unexpected occurrences of events. Since the patterns that event sequences tend to follow may change in different contexts, we develop outlier detection methods based on point processes that can take context information into account. Our methods are based on Bayesian decision theory and hypothesis testing with theoretical guarantees. To test the performance of the methods, we conduct experiments on both synthetic data and real-world clinical data and show the effectiveness of the proposed methods.
****
Heterogeneous Risk Minimization
Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, Zheyan Shen
Machine learning algorithms with empirical risk minimization usually suffer from poor generalization performance due to the greedy exploitation of correlations among the training data, which are not stable under distributional shifts. Recently, some invariant learning methods for out-of-distribution (OOD) generalization have been proposed by leveraging multiple training environments to find invariant relationships. However, modern datasets are frequently assembled by merging data from multiple sources without explicit source labels. The resultant unobserved heterogeneity renders many invariant learning methods inapplicable. In this paper, we propose Heterogeneous Risk Minimization (HRM) framework to achieve joint learning of latent heterogeneity among the data and invariant relationship, which leads to stable prediction despite distributional shifts. We theoretically characterize the roles of the environment labels in invariant learning and justify our newly proposed HRM framework. Extensive experimental results validate the effectiveness of our HRM framework.
****
Stochastic Iterative Graph Matching
Linfeng Liu, Michael C Hughes, Soha Hassoun, Liping Liu
Recent works apply Graph Neural Networks (GNNs) to graph matching tasks and show promising results. Considering that model outputs are complex matchings, we devise several techniques to improve the learning of GNNs and obtain a new model, Stochastic Iterative Graph MAtching (SIGMA). Our model predicts a distribution of matchings, instead of a single matching, for a graph pair so the model can explore several probable matchings. We further introduce a novel multi-step matching procedure, which learns how to refine a graph pair’s matching results increment

ally. The model also includes dummy nodes so that the model does not have to find matchings for nodes without correspondence. We fit this model to data via scalable stochastic optimization. We conduct extensive experiments across synthetic graph datasets as well as biochemistry and computer vision applications. Across all tasks, our results show that SIGMA can produce significantly improved graph matching results compared to state-of-the-art models. Ablation studies verify that each of our components (stochastic training, iterative matching, and dummy nodes) offers noticeable improvement.
****
Cooperative Exploration for Multi-Agent Deep Reinforcement Learning
Iou-Jen Liu, Unnat Jain, Raymond A Yeh, Alexander Schwing
Exploration is critical for good results in deep reinforcement learning and has attracted much attention. However, existing multi-agent deep reinforcement learning algorithms still use mostly noise-based techniques. Very recently, exploration methods that consider cooperation among multiple agents have been developed. However, existing methods suffer from a common challenge: agents struggle to identify states that are worth exploring, and hardly coordinate exploration efforts toward those states. To address this shortcoming, in this paper, we propose cooperative multi-agent exploration (CMAE): agents share a common goal while exploring. The goal is selected from multiple projected state spaces by a normalized entropy-based technique. Then, agents are trained to reach the goal in a coordinated manner. We demonstrate that CMAE consistently outperforms baselines on various tasks, including a sparse-reward version of multiple-particle environment (MPE) and the Starcraft multi-agent challenge (SMAC).
****
Elastic Graph Neural Networks
Xiaorui Liu, Wei Jin, Yao Ma, Yaxin Li, Hua Liu, Yiqi Wang, Ming Yan, Jiliang Tang
While many existing graph neural networks (GNNs) have been proven to perform ℓ_2-based graph smoothing that enforces smoothness globally, in this work we aim to further enhance the local smoothness adaptivity of GNNs via ℓ_1-based graph smoothing. As a result, we introduce a family of GNNs (Elastic GNNs) based on ℓ_1 and ℓ_2-based graph smoothing. In particular, we propose a novel and general message passing scheme into GNNs. This message passing algorithm is not only friendly to back-propagation training but also achieves the desired smoothing properties with a theoretical convergence guarantee. Experiments on semi-supervised learning tasks demonstrate that the proposed Elastic GNNs obtain better adaptivity on benchmark datasets and are significantly robust to graph adversarial attacks. The implementation of Elastic GNNs is available at \url{https://github.com/lxiaorui/ElasticGNN}.
****
One Pass Late Fusion Multi-view Clustering
Xinwang Liu, Li Liu, Qing Liao, Siwei Wang, Yi Zhang, Wenxuan Tu, Chang Tang, Jiyuan Liu, En Zhu
Existing late fusion multi-view clustering (LFMVC) optimally integrates a group of pre-specified base partition matrices to learn a consensus one. It is then taken as the input of the widely used k-means to generate the cluster labels. As observed, the learning of the consensus partition matrix and the generation of cluster labels are separately done. These two procedures lack necessary negotiation and can not best serve for each other, which may adversely affect the clustering performance. To address this issue, we propose to unify the aforementioned two learning procedures into a single optimization, in which the consensus partition matrix can better serve for the generation of cluster labels, and the latter is able to guide the learning of the former. To optimize the resultant optimization problem, we develop a four-step alternate algorithm with proved convergence. We theoretically analyze the clustering generalization error of the proposed algorithm on unseen data. Comprehensive experiments on multiple benchmark datasets demonstrate the superiority of our algorithm in terms of both clustering accuracy and computational efficiency. It is expected that the simplicity and effectiveness of our algorithm will make it a good option to be considered for practical

multi-view clustering applications.
****
Coach-Player Multi-agent Reinforcement Learning for Dynamic Team Composition
Bo Liu, Qiang Liu, Peter Stone, Animesh Garg, Yuke Zhu, Anima Anandkumar
In real-world multi-agent systems, agents with different capabilities may join o
r leave without altering the team's overarching goals. Coordinating teams with s
uch dynamic composition is challenging: the optimal team strategy varies with th
e composition. We propose COPA, a coach-player framework to tackle this problem.
 We assume the coach has a global view of the environment and coordinates the pl
ayers, who only have partial views, by distributing individual strategies. Speci
fically, we 1) adopt the attention mechanism for both the coach and the players;
 2) propose a variational objective to regularize learning; and 3) design an ada
ptive communication method to let the coach decide when to communicate with the
players. We validate our methods on a resource collection task, a rescue game, a
nd the StarCraft micromanagement tasks. We demonstrate zero-shot generalization
to new team compositions. Our method achieves comparable or better performance t
han the setting where all players have a full view of the environment. Moreover,
 we see that the performance remains high even when the coach communicates as li
ttle as 13% of the time using the adaptive communication strategy.
****
From Local to Global Norm Emergence: Dissolving Self-reinforcing Substructures w
ith Incremental Social Instruments
Yiwei Liu, Jiamou Liu, Kaibin Wan, Zhan Qin, Zijian Zhang, Bakhadyr Khoussainov,
 Liehuang Zhu
Norm emergence is a process where agents in a multi-agent system establish self-
enforcing conformity through repeated interactions. When such interactions are c
onfined to a social topology, several self-reinforcing substructures (SRS) may e
merge within the population. This prevents a formation of a global norm. We prop
ose incremental social instruments (ISI) to dissolve these SRSs by creating ties
 between agents. Establishing ties requires some effort and cost. Hence, it is w
orth to design methods that build a small number of ties yet dissolve the SRSs.
By using the notion of information entropy, we propose an indicator called the B
A-ratio that measures the current SRSs. We find that by building ties with minim
al BA-ratio, our ISI is effective in facilitating the global norm emergence. We
explain this through our experiments and theoretical results. Furthermore, we pr
opose the small-degree principle in minimising the BA-ratio that helps us to des
ign efficient ISI algorithms for finding the optimal ties. Experiments on both s
ynthetic and real-world network topologies demonstrate that our adaptive ISI is
efficient at dissolving SRS.
****
A Value-Function-based Interior-point Method for Non-convex Bi-level Optimizatio
n
Risheng Liu, Xuan Liu, Xiaoming Yuan, Shangzhi Zeng, Jin Zhang
Bi-level optimization model is able to capture a wide range of complex learning
tasks with practical interest. Due to the witnessed efficiency in solving bi-lev
el programs, gradient-based methods have gained popularity in the machine learni
ng community. In this work, we propose a new gradient-based solution scheme, nam
ely, the Bi-level Value-Function-based Interior-point Method (BVFIM). Following
the main idea of the log-barrier interior-point scheme, we penalize the regulari
zed value function of the lower level problem into the upper level objective. By
 further solving a sequence of differentiable unconstrained approximation proble
ms, we consequently derive a sequential programming scheme. The numerical advant
age of our scheme relies on the fact that, when gradient methods are applied to
solve the approximation problem, we successfully avoid computing any expensive H
essian-vector or Jacobian-vector product. We prove the convergence without requi
ring any convexity assumption on either the upper level or the lower level objec
tive. Experiments demonstrate the efficiency of the proposed BVFIM on non-convex
 bi-level problems.
****
Selfish Sparse RNN Training

Shiwei Liu, Decebal Constantin Mocanu, Yulong Pei, Mykola Pechenizkiy
Sparse neural networks have been widely applied to reduce the computational demands of training and deploying over-parameterized deep neural networks. For inference acceleration, methods that discover a sparse network from a pre-trained dense network (dense-to-sparse training) work effectively. Recently, dynamic sparse training (DST) has been proposed to train sparse neural networks without pre-training a dense model (sparse-to-sparse training), so that the training process can also be accelerated. However, previous sparse-to-sparse methods mainly focus on Multilayer Perceptron Networks (MLPs) and Convolutional Neural Networks (CNNs), failing to match the performance of dense-to-sparse methods in the Recurrent Neural Networks (RNNs) setting. In this paper, we propose an approach to train intrinsically sparse RNNs with a fixed parameter count in one single run, without compromising performance. During training, we allow RNN layers to have a non-uniform redistribution across cell gates for better regularization. Further, we propose SNT-ASGD, a novel variant of the averaged stochastic gradient optimizer, which significantly improves the performance of all sparse training methods for RNNs. Using these strategies, we achieve state-of-the-art sparse training results, better than the dense-to-sparse methods, with various types of RNNs on Penn TreeBank and Wikitext-2 datasets. Our codes are available at https://github.com/Shiweiliuiiiiiii/Selfish-RNN.
****
Temporal Difference Learning as Gradient Splitting
Rui Liu, Alex Olshevsky
Temporal difference learning with linear function approximation is a popular method to obtain a low-dimensional approximation of the value function of a policy in a Markov Decision Process. We provide an interpretation of this method in terms of a splitting of the gradient of an appropriately chosen function. As a consequence of this interpretation, convergence proofs for gradient descent can be applied almost verbatim to temporal difference learning. Beyond giving a fuller explanation of why temporal difference works, this interpretation also yields improved convergence times. We consider the setting with $1/\sqrt{T}$ step-size, where previous comparable finite-time convergence time bounds for temporal difference learning had the multiplicative factor $1/(1-\gamma)$ in front of the bound, with γ being the discount factor. We show that a minor variation on TD learning which estimates the mean of the value function separately has a convergence time where $1/(1-\gamma)$ only multiplies an asymptotically negligible term.
****
On Robust Mean Estimation under Coordinate-level Corruption
Zifan Liu, Jong Ho Park, Theodoros Rekatsinas, Christos Tzamos
We study the problem of robust mean estimation and introduce a novel Hamming distance-based measure of distribution shift for coordinate-level corruptions. We show that this measure yields adversary models that capture more realistic corruptions than those used in prior works, and present an information-theoretic analysis of robust mean estimation in these settings. We show that for structured distributions, methods that leverage the structure yield information theoretically more accurate mean estimation. We also focus on practical algorithms for robust mean estimation and study when data cleaning-inspired approaches that first fix corruptions in the input data and then perform robust mean estimation can match the information theoretic bounds of our analysis. We finally demonstrate experimentally that this two-step approach outperforms structure-agnostic robust estimation and provides accurate mean estimation even for high-magnitude corruption.
****
Decoupling Exploration and Exploitation for Meta-Reinforcement Learning without Sacrifices
Evan Z Liu, Aditi Raghunathan, Percy Liang, Chelsea Finn
The goal of meta-reinforcement learning (meta-RL) is to build agents that can quickly learn new tasks by leveraging prior experience on related tasks. Learning a new task often requires both exploring to gather task-relevant information and exploiting this information to solve the task. In principle, optimal exploration and exploitation can be learned end-to-end by simply maximizing task performan

ce. However, such meta-RL approaches struggle with local optima due to a chicken-and-egg problem: learning to explore requires good exploitation to gauge the exploration’s utility, but learning to exploit requires information gathered via exploration. Optimizing separate objectives for exploration and exploitation can avoid this problem, but prior meta-RL exploration objectives yield suboptimal policies that gather information irrelevant to the task. We alleviate both concerns by constructing an exploitation objective that automatically identifies task-relevant information and an exploration objective to recover only this information. This avoids local optima in end-to-end training, without sacrificing optimal exploration. Empirically, DREAM substantially outperforms existing approaches on complex meta-RL problems, such as sparse-reward 3D visual navigation. Videos of DREAM: https://ezliu.github.io/dream/
****
How Do Adam and Training Strategies Help BNNs Optimization
Zechun Liu, Zhiqiang Shen, Shichao Li, Koen Helwegen, Dong Huang, Kwang-Ting Cheng
The best performing Binary Neural Networks (BNNs) are usually attained using Adam optimization and its multi-step training variants. However, to the best of our knowledge, few studies explore the fundamental reasons why Adam is superior to other optimizers like SGD for BNN optimization or provide analytical explanations that support specific training strategies. To address this, in this paper we first investigate the trajectories of gradients and weights in BNNs during the training process. We show the regularization effect of second-order momentum in Adam is crucial to revitalize the weights that are dead due to the activation saturation in BNNs. We find that Adam, through its adaptive learning rate strategy, is better equipped to handle the rugged loss surface of BNNs and reaches a better optimum with higher generalization ability. Furthermore, we inspect the intriguing role of the real-valued weights in binary networks, and reveal the effect of weight decay on the stability and sluggishness of BNN optimization. Through extensive experiments and analysis, we derive a simple training scheme, building on existing Adam-based optimization, which achieves 70.5% top-1 accuracy on the ImageNet dataset using the same architecture as the state-of-the-art ReActNet while achieving 1.1% higher accuracy. Code and models are available at https://github.com/liuzechun/AdamBNN.
****
SagaNet: A Small Sample Gated Network for Pediatric Cancer Diagnosis
Yuhan Liu, Shiliang Sun
The scarcity of available samples and the high annotation cost of medical data cause a bottleneck in many digital diagnosis tasks based on deep learning. This problem is especially severe in pediatric tumor tasks, due to the small population base of children and high sample diversity caused by the high metastasis rate of related tumors. Targeted research on pediatric tumors is urgently needed but lacks sufficient attention. In this work, we propose a novel model to solve the diagnosis task of small round blue cell tumors (SRBCTs). To solve the problem of high noise and high diversity in the small sample scenario, the model is constrained to pay attention to the valid areas in the pathological image with a masking mechanism, and a length-aware loss is proposed to improve the tolerance to feature diversity. We evaluate this framework on a challenging small sample SRBCTs dataset, whose classification is difficult even for professional pathologists. The proposed model shows the best performance compared with state-of-the-art deep models and generalization on another pathological dataset, which illustrates the potentiality of deep learning applications in difficult small sample medical tasks.
****
Learning Deep Neural Networks under Agnostic Corrupted Supervision
Boyang Liu, Mengying Sun, Ding Wang, Pang-Ning Tan, Jiayu Zhou
Training deep neural network models in the presence of corrupted supervision is challenging as the corrupted data points may significantly impact generalization performance. To alleviate this problem, we present an efficient robust algorithm that achieves strong guarantees without any assumption on the type of corrupti

on and provides a unified framework for both classification and regression problems. Unlike many existing approaches that quantify the quality of the data points (e.g., based on their individual loss values), and filter them accordingly, the proposed algorithm focuses on controlling the collective impact of data points on the average gradient. Even when a corrupted data point failed to be excluded by our algorithm, the data point will have a very limited impact on the overall loss, as compared with state-of-the-art filtering methods based on loss values. Extensive experiments on multiple benchmark datasets have demonstrated the robustness of our algorithm under different types of corruption. Our code is available at \url{https://github.com/illidanlab/PRL}.
****
Leveraging Public Data for Practical Private Query Release
Terrance Liu, Giuseppe Vietri, Thomas Steinke, Jonathan Ullman, Steven Wu
In many statistical problems, incorporating priors can significantly improve performance. However, the use of prior knowledge in differentially private query release has remained underexplored, despite such priors commonly being available in the form of public datasets, such as previous US Census releases. With the goal of releasing statistics about a private dataset, we present PMW^Pub, which—unlike existing baselines—leverages public data drawn from a related distribution as prior information. We provide a theoretical analysis and an empirical evaluation on the American Community Survey (ACS) and ADULT datasets, which shows that our method outperforms state-of-the-art methods. Furthermore, PMW^Pub scales well to high-dimensional data domains, where running many existing methods would be computationally infeasible.
****
Watermarking Deep Neural Networks with Greedy Residuals
Hanwen Liu, Zhenyu Weng, Yuesheng Zhu
Deep neural networks (DNNs) are considered as intellectual property of their corresponding owners and thus are in urgent need of ownership protection, due to the massive amount of time and resources invested in designing, tuning and training them. In this paper, we propose a novel watermark-based ownership protection method by using the residuals of important parameters. Different from other watermark-based ownership protection methods that rely on some specific neural network architectures and during verification require external data source, namely ownership indicators, our method does not explicitly use ownership indicators for verification to defeat various attacks against DNN watermarks. Specifically, we greedily select a few and important model parameters for embedding so that the impairment caused by the changed parameters can be reduced and the robustness against different attacks can be improved as the selected parameters can well preserve the model information. Also, without the external data sources for verification, the adversary can hardly cast doubts on ownership verification by forging counterfeit watermarks. The extensive experiments show that our method outperforms previous state-of-the-art methods in five tasks.
****
Do We Actually Need Dense Over-Parameterization? In-Time Over-Parameterization in Sparse Training
Shiwei Liu, Lu Yin, Decebal Constantin Mocanu, Mykola Pechenizkiy
In this paper, we introduce a new perspective on training deep neural networks capable of state-of-the-art performance without the need for the expensive over-parameterization by proposing the concept of In-Time Over-Parameterization (ITOP) in sparse training. By starting from a random sparse network and continuously exploring sparse connectivities during training, we can perform an Over-Parameterization over the course of training, closing the gap in the expressibility between sparse training and dense training. We further use ITOP to understand the underlying mechanism of Dynamic Sparse Training (DST) and discover that the benefits of DST come from its ability to consider across time all possible parameters when searching for the optimal sparse connectivity. As long as sufficient parameters have been reliably explored, DST can outperform the dense neural network by a large margin. We present a series of experiments to support our conjecture and achieve the state-of-the-art sparse training performance with ResNet-50 on Imag

eNet. More impressively, ITOP achieves dominant performance over the overparameterization-based sparse methods at extreme sparsities. When trained with ResNet-34 on CIFAR-100, ITOP can match the performance of the dense model at an extreme sparsity 98%.
****
A Sharp Analysis of Model-based Reinforcement Learning with Self-Play
Qinghua Liu, Tiancheng Yu, Yu Bai, Chi Jin
Model-based algorithms—algorithms that explore the environment through building and utilizing an estimated model—are widely used in reinforcement learning practice and theoretically shown to achieve optimal sample efficiency for single-agent reinforcement learning in Markov Decision Processes (MDPs). However, for multi-agent reinforcement learning in Markov games, the current best known sample complexity for model-based algorithms is rather suboptimal and compares unfavorably against recent model-free approaches. In this paper, we present a sharp analysis of model-based self-play algorithms for multi-agent Markov games. We design an algorithm \emph{Optimistic Nash Value Iteration} (Nash-VI) for two-player zero-sum Markov games that is able to output an ϵ-approximate Nash policy in $\tilde{\mathcal{O}}(H^3SAB/\epsilon^2)$ episodes of game playing, where S is the number of states, A,B are the number of actions for the two players respectively, and H is the horizon length. This significantly improves over the best known model-based guarantee of $\tilde{\mathcal{O}}(H^4S^2AB/\epsilon^2)$, and is the first that matches the information-theoretic lower bound $\Omega(H^3S(A+B)/\epsilon^2)$ except for a $\min\{A,B\}$ factor. In addition, our guarantee compares favorably against the best known model-free algorithm if $\min\{A,B\}=o(H^3)$, and outputs a single Markov policy while existing sample-efficient model-free algorithms output a nested mixture of Markov policies that is in general non-Markov and rather inconvenient to store and execute. We further adapt our analysis to designing a provably efficient task-agnostic algorithm for zero-sum Markov games, and designing the first line of provably sample-efficient algorithms for multi-player general-sum Markov games.
****
Lottery Ticket Preserves Weight Correlation: Is It Desirable or Not?
Ning Liu, Geng Yuan, Zhengping Che, Xuan Shen, Xiaolong Ma, Qing Jin, Jian Ren, Jian Tang, Sijia Liu, Yanzhi Wang
In deep model compression, the recent finding "Lottery Ticket Hypothesis" (LTH) pointed out that there could exist a winning ticket (i.e., a properly pruned sub-network together with original weight initialization) that can achieve competitive performance than the original dense network. However, it is not easy to observe such winning property in many scenarios, where for example, a relatively large learning rate is used even if it benefits training the original dense model. In this work, we investigate the underlying condition and rationale behind the winning property, and find that the underlying reason is largely attributed to the correlation between initialized weights and final-trained weights when the learning rate is not sufficiently large. Thus, the existence of winning property is correlated with an insufficient DNN pretraining, and is unlikely to occur for a well-trained DNN. To overcome this limitation, we propose the "pruning & fine-tuning" method that consistently outperforms lottery ticket sparse training under the same pruning algorithm and the same total training epochs. Extensive experiments over multiple deep models (VGG, ResNet, MobileNet-v2) on different datasets have been conducted to justify our proposals.
****
Group Fisher Pruning for Practical Network Compression
Liyang Liu, Shilong Zhang, Zhanghui Kuang, Aojun Zhou, Jing-Hao Xue, Xinjiang Wang, Yimin Chen, Wenming Yang, Qingmin Liao, Wayne Zhang
Network compression has been widely studied since it is able to reduce the memory and computation cost during inference. However, previous methods seldom deal with complicated structures like residual connections, group/depth-wise convolution and feature pyramid network, where channels of multiple layers are coupled and need to be pruned simultaneously. In this paper, we present a general channel pruning approach that can be applied to various complicated structures. Particul

arly, we propose a layer grouping algorithm to find coupled channels automatically. Then we derive a unified metric based on Fisher information to evaluate the importance of a single channel and coupled channels. Moreover, we find that inference speedup on GPUs is more correlated with the reduction of memory rather than FLOPs, and thus we employ the memory reduction of each channel to normalize the importance. Our method can be used to prune any structures including those with coupled channels. We conduct extensive experiments on various backbones, including the classic ResNet and ResNeXt, mobile-friendly MobileNetV2, and the NAS-based RegNet, both on image classification and object detection which is under-explored. Experimental results validate that our method can effectively prune sophisticated networks, boosting inference speed without sacrificing accuracy.
****
Infinite-Dimensional Optimization for Zero-Sum Games via Variational Transport
Lewis Liu, Yufeng Zhang, Zhuoran Yang, Reza Babanezhad, Zhaoran Wang
Game optimization has been extensively studied when decision variables lie in a finite-dimensional space, of which solutions correspond to pure strategies at the Nash equilibrium (NE), and the gradient descent-ascent (GDA) method works widely in practice. In this paper, we consider infinite-dimensional zero-sum games by a min-max distributional optimization problem over a space of probability measures defined on a continuous variable set, which is inspired by finding a mixed NE for finite-dimensional zero-sum games. We then aim to answer the following question: \textit{Will GDA-type algorithms still be provably efficient when extended to infinite-dimensional zero-sum games?} To answer this question, we propose a particle-based variational transport algorithm based on GDA in the functional spaces. Specifically, the algorithm performs multi-step functional gradient descent-ascent in the Wasserstein space via pushing two sets of particles in the variable space. By characterizing the gradient estimation error from variational form maximization and the convergence behavior of each player with different objective landscapes, we prove rigorously that the generalized GDA algorithm converges to the NE or the value of the game efficiently for a class of games under the Polyak-■{ojasiewicz} (PL) condition. To conclude, we provide complete statistical and convergence guarantees for solving an infinite-dimensional zero-sum game via a provably efficient particle-based method. Additionally, our work provides the first thorough statistical analysis for the particle-based algorithm to learn an objective functional with a variational form using universal approximators (\textit{i.e.}, neural networks (NNs)), which is of independent interest.
****
Noise and Fluctuation of Finite Learning Rate Stochastic Gradient Descent
Kangqiao Liu, Liu Ziyin, Masahito Ueda
In the vanishing learning rate regime, stochastic gradient descent (SGD) is now relatively well understood. In this work, we propose to study the basic properties of SGD and its variants in the non-vanishing learning rate regime. The focus is on deriving exactly solvable results and discussing their implications. The main contributions of this work are to derive the stationary distribution for discrete-time SGD in a quadratic loss function with and without momentum; in particular, one implication of our result is that the fluctuation caused by discrete-time dynamics takes a distorted shape and is dramatically larger than a continuous-time theory could predict. Examples of applications of the proposed theory considered in this work include the approximation error of variants of SGD, the effect of minibatch noise, the optimal Bayesian inference, the escape rate from a sharp minimum, and the stationary covariance of a few second-order methods including damped Newton’s method, natural gradient descent, and Adam.
****
Multi-layered Network Exploration via Random Walks: From Offline Optimization to Online Learning
Xutong Liu, Jinhang Zuo, Xiaowei Chen, Wei Chen, John C. S. Lui
Multi-layered network exploration (MuLaNE) problem is an important problem abstracted from many applications. In MuLaNE, there are multiple network layers where each node has an importance weight and each layer is explored by a random walk. The MuLaNE task is to allocate total random walk budget B into each network l

ayer so that the total weights of the unique nodes visited by random walks are m
aximized. We systematically study this problem from offline optimization to onli
ne learning. For the offline optimization setting where the network structure an
d node weights are known, we provide greedy based constant-ratio approximation a
lgorithms for overlapping networks, and greedy or dynamic-programming based opti
mal solutions for non-overlapping networks. For the online learning setting, nei
ther the network structure nor the node weights are known initially. We adapt th
e combinatorial multi-armed bandit framework and design algorithms to learn rand
om walk related parameters and node weights while optimizing the budget allocati
on in multiple rounds, and prove that they achieve logarithmic regret bounds. Fi
nally, we conduct experiments on a real-world social network dataset to validate
 our theoretical results.
****
Relative Positional Encoding for Transformers with Linear Complexity
Antoine Liutkus, Ond■ej C■■fka, Shih-Lun Wu, Umut Simsekli, Yi-Hsuan Yang, Gael
Richard
Recent advances in Transformer models allow for unprecedented sequence lengths,
due to linear space and time complexity. In the meantime, relative positional en
coding (RPE) was proposed as beneficial for classical Transformers and consists
in exploiting lags instead of absolute positions for inference. Still, RPE is no
t available for the recent linear-variants of the Transformer, because it requir
es the explicit computation of the attention matrix, which is precisely what is
avoided by such methods. In this paper, we bridge this gap and present Stochasti
c Positional Encoding as a way to generate PE that can be used as a replacement
to the classical additive (sinusoidal) PE and provably behaves like RPE. The mai
n theoretical contribution is to make a connection between positional encoding a
nd cross-covariance structures of correlated Gaussian processes. We illustrate t
he performance of our approach on the Long-Range Arena benchmark and on music ge
neration.
****
Joint Online Learning and Decision-making via Dual Mirror Descent
Alfonso Lobos, Paul Grigas, Zheng Wen
We consider an online revenue maximization problem over a finite time horizon su
bject to lower and upper bounds on cost. At each period, an agent receives a con
text vector sampled i.i.d. from an unknown distribution and needs to make a deci
sion adaptively. The revenue and cost functions depend on the context vector as
well as some fixed but possibly unknown parameter vector to be learned. We propo
se a novel offline benchmark and a new algorithm that mixes an online dual mirro
r descent scheme with a generic parameter learning process. When the parameter v
ector is known, we demonstrate an $O(\sqrt{T})$ regret result as well an $O(\sqr
t{T})$ bound on the possible constraint violations. When the parameter is not kn
own and must be learned, we demonstrate that the regret and constraint violation
s are the sums of the previous $O(\sqrt{T})$ terms plus terms that directly depe
nd on the convergence of the learning process.
****
Symmetric Spaces for Graph Embeddings: A Finsler-Riemannian Approach
Federico Lopez, Beatrice Pozzetti, Steve Trettel, Michael Strube, Anna Wienhard
Learning faithful graph representations as sets of vertex embeddings has become
a fundamental intermediary step in a wide range of machine learning applications
. We propose the systematic use of symmetric spaces in representation learning,
a class encompassing many of the previously used embedding targets. This enables
 us to introduce a new method, the use of Finsler metrics integrated in a Rieman
nian optimization scheme, that better adapts to dissimilar structures in the gra
ph. We develop a tool to analyze the embeddings and infer structural properties
of the data sets. For implementation, we choose Siegel spaces, a versatile famil
y of symmetric spaces. Our approach outperforms competitive baselines for graph
reconstruction tasks on various synthetic and real-world datasets. We further de
monstrate its applicability on two downstream tasks, recommender systems and nod
e classification.
****

HEMET: A Homomorphic-Encryption-Friendly Privacy-Preserving Mobile Neural Network Architecture
Qian Lou, Lei Jiang
Recently Homomorphic Encryption (HE) is used to implement Privacy-Preserving Neural Networks (PPNNs) that perform inferences directly on encrypted data without decryption. Prior PPNNs adopt mobile network architectures such as SqueezeNet for smaller computing overhead, but we find naïvely using mobile network architectures for a PPNN does not necessarily achieve shorter inference latency. Despite having less parameters, a mobile network architecture typically introduces more layers and increases the HE multiplicative depth of a PPNN, thereby prolonging its inference latency. In this paper, we propose a \textbf{HE}-friendly privacy-preserving \textbf{M}obile neural n\textbf{ET}work architecture, \textbf{HEMET}. Experimental results show that, compared to state-of-the-art (SOTA) PPNNs, HEMET reduces the inference latency by $59.3%\sim 61.2%$, and improves the inference accuracy by $0.4 % \sim 0.5%$.
****
Optimal Complexity in Decentralized Training
Yucheng Lu, Christopher De Sa
Decentralization is a promising method of scaling up parallel machine learning systems. In this paper, we provide a tight lower bound on the iteration complexity for such methods in a stochastic non-convex setting. Our lower bound reveals a theoretical gap in known convergence rates of many existing decentralized training algorithms, such as D-PSGD. We prove by construction this lower bound is tight and achievable. Motivated by our insights, we further propose DeTAG, a practical gossip-style decentralized algorithm that achieves the lower bound with only a logarithm gap. Empirically, we compare DeTAG with other decentralized algorithms on image classification tasks, and we show DeTAG enjoys faster convergence compared to baselines, especially on unshuffled data and in sparse networks.
****
DANCE: Enhancing saliency maps using decoys
Yang Young Lu, Wenbo Guo, Xinyu Xing, William Stafford Noble
Saliency methods can make deep neural network predictions more interpretable by identifying a set of critical features in an input sample, such as pixels that contribute most strongly to a prediction made by an image classifier. Unfortunately, recent evidence suggests that many saliency methods poorly perform, especially in situations where gradients are saturated, inputs contain adversarial perturbations, or predictions rely upon inter-feature dependence. To address these issues, we propose a framework, DANCE, which improves the robustness of saliency methods by following a two-step procedure. First, we introduce a perturbation mechanism that subtly varies the input sample without changing its intermediate representations. Using this approach, we can gather a corpus of perturbed ("decoy") data samples while ensuring that the perturbed and original input samples follow similar distributions. Second, we compute saliency maps for the decoy samples and propose a new method to aggregate saliency maps. With this design, we offset influence of gradient saturation. From a theoretical perspective, we show that the aggregated saliency map not only captures inter-feature dependence but, more importantly, is robust against previously described adversarial perturbation methods. Our empirical results suggest that, both qualitatively and quantitatively, DANCE outperforms existing methods in a variety of application domains.
****
Binary Classification from Multiple Unlabeled Datasets via Surrogate Set Classification
Nan Lu, Shida Lei, Gang Niu, Issei Sato, Masashi Sugiyama
To cope with high annotation costs, training a classifier only from weakly supervised data has attracted a great deal of attention these days. Among various approaches, strengthening supervision from completely unsupervised classification is a promising direction, which typically employs class priors as the only supervision and trains a binary classifier from unlabeled (U) datasets. While existing risk-consistent methods are theoretically grounded with high flexibility, they can learn only from two U sets. In this paper, we propose a new approach for bin

ary classification from m U-sets for $m\ge2$. Our key idea is to consider an a uxiliary classification task called surrogate set classification (SSC), which is aimed at predicting from which U set each observed sample is drawn. SSC can be solved by a standard (multi-class) classification method, and we use the SSC sol ution to obtain the final binary classifier through a certain linear-fractional transformation. We built our method in a flexible and efficient end-to-end deep learning framework and prove it to be classifier-consistent. Through experiments , we demonstrate the superiority of our proposed method over state-of-the-art me thods.
****
Variance Reduced Training with Stratified Sampling for Forecasting Models
Yucheng Lu, Youngsuk Park, Lifan Chen, Yuyang Wang, Christopher De Sa, Dean Fost er
In large-scale time series forecasting, one often encounters the situation where the temporal patterns of time series, while drifting over time, differ from one another in the same dataset. In this paper, we provably show under such heterog eneity, training a forecasting model with commonly used stochastic optimizers (e .g. SGD) potentially suffers large variance on gradient estimation, and thus inc urs long-time training. We show that this issue can be efficiently alleviated vi a stratification, which allows the optimizer to sample from pre-grouped time ser ies strata. For better trading-off gradient variance and computation complexity, we further propose SCott (Stochastic Stratified Control Variate Gradient Descen t), a variance reduced SGD-style optimizer that utilizes stratified sampling via control variate. In theory, we provide the convergence guarantee of SCott on sm ooth non-convex objectives. Empirically, we evaluate SCott and other baseline op timizers on both synthetic and real-world time series forecasting problems, and demonstrate SCott converges faster with respect to both iterations and wall cloc k time.
****
ACE: Explaining cluster from an adversarial perspective
Yang Young Lu, Timothy C Yu, Giancarlo Bonora, William Stafford Noble
A common workflow in single-cell RNA-seq analysis is to project the data to a la tent space, cluster the cells in that space, and identify sets of marker genes t hat explain the differences among the discovered clusters. A primary drawback to this three-step procedure is that each step is carried out independently, there by neglecting the effects of the nonlinear embedding and inter-gene dependencies on the selection of marker genes. Here we propose an integrated deep learning f ramework, Adversarial Clustering Explanation (ACE), that bundles all three steps into a single workflow. The method thus moves away from the notion of "marker g enes" to instead identify a panel of explanatory genes. This panel may include g enes that are not only enriched but also depleted relative to other cell types, as well as genes that exhibit differences between closely related cell types. Em pirically, we demonstrate that ACE is able to identify gene panels that are both highly discriminative and nonredundant, and we demonstrate the applicability of ACE to an image recognition task.
****
On Monotonic Linear Interpolation of Neural Network Parameters
James R Lucas, Juhan Bae, Michael R Zhang, Stanislav Fort, Richard Zemel, Roger B Grosse
Linear interpolation between initial neural network parameters and converged par ameters after training with stochastic gradient descent (SGD) typically leads to a monotonic decrease in the training objective. This Monotonic Linear Interpola tion (MLI) property, first observed by Goodfellow et al. 2014, persists in spite of the non-convex objectives and highly non-linear training dynamics of neural networks. Extending this work, we evaluate several hypotheses for this property that, to our knowledge, have not yet been explored. Using tools from differentia l geometry, we draw connections between the interpolated paths in function space and the monotonicity of the network — providing sufficient conditions for the M LI property under mean squared error. While the MLI property holds under various settings (e.g., network architectures and learning problems), we show in practi

ce that networks violating the MLI property can be produced systematically, by e
ncouraging the weights to move far from initialization. The MLI property raises
important questions about the loss landscape geometry of neural networks and hig
hlights the need to further study their global properties.
****
Improving Breadth-Wise Backpropagation in Graph Neural Networks Helps Learning L
ong-Range Dependencies.
Denis Lukovnikov, Asja Fischer
In this work, we focus on the ability of graph neural networks (GNNs) to learn l
ong-range patterns in graphs with edge features. Learning patterns that involve
longer paths in the graph, requires using deeper GNNs. However, GNNs suffer from
 a drop in performance with increasing network depth. To improve the performance
 of deeper GNNs, previous works have investigated normalization techniques and v
arious types of skip connections. While they are designed to improve depth-wise
backpropagation between the representations of the same node in successive layer
s, they do not improve breadth-wise backpropagation between representations of n
eighbouring nodes. To analyse the consequences, we design synthetic datasets ser
ving as a testbed for the ability of GNNs to learn long-range patterns. Our anal
ysis shows that several commonly used GNN variants with only depth-wise skip con
nections indeed have problems learning long-range patterns. They are clearly out
performed by an attention-based GNN architecture that we propose for improving b
oth depth- and breadth-wise backpropagation. We also verify that the presented a
rchitecture is competitive on real-world data.
****
GraphDF: A Discrete Flow Model for Molecular Graph Generation
Youzhi Luo, Keqiang Yan, Shuiwang Ji
We consider the problem of molecular graph generation using deep models. While g
raphs are discrete, most existing methods use continuous latent variables, resul
ting in inaccurate modeling of discrete graph structures. In this work, we propo
se GraphDF, a novel discrete latent variable model for molecular graph generatio
n based on normalizing flow methods. GraphDF uses invertible modulo shift transf
orms to map discrete latent variables to graph nodes and edges. We show that the
 use of discrete latent variables reduces computational costs and eliminates the
 negative effect of dequantization. Comprehensive experimental results show that
 GraphDF outperforms prior methods on random generation, property optimization,
and constrained optimization tasks.
****
Trajectory Diversity for Zero-Shot Coordination
Andrei Lupu, Brandon Cui, Hengyuan Hu, Jakob Foerster
We study the problem of zero-shot coordination (ZSC), where agents must independ
ently produce strategies for a collaborative game that are compatible with novel
 partners not seen during training. Our first contribution is to consider the ne
ed for diversity in generating such agents. Because self-play (SP) agents contro
l their own trajectory distribution during training, each policy typically only
performs well on this exact distribution. As a result, they achieve low scores i
n ZSC, since playing with another agent is likely to put them in situations they
 have not encountered during training. To address this issue, we train a common
best response (BR) to a population of agents, which we regulate to be diverse. T
o this end, we introduce \textit{Trajectory Diversity} (TrajeDi) – a differentia
ble objective for generating diverse reinforcement learning policies. We derive
TrajeDi as a generalization of the Jensen-Shannon divergence between policies an
d motivate it experimentally in two simple settings. We then focus on the collab
orative card game Hanabi, demonstrating the scalability of our method and improv
ing upon the cross-play scores of both independently trained SP agents and BRs t
o unregularized populations.
****
HyperHyperNetwork for the Design of Antenna Arrays
Shahar Lutati, Lior Wolf
We present deep learning methods for the design of arrays and single instances o
f small antennas. Each design instance is conditioned on a target radiation patt

ern and is required to conform to specific spatial dimensions and to include, as part of its metallic structure, a set of predetermined locations. The solution, in the case of a single antenna, is based on a composite neural network that combines a simulation network, a hypernetwork, and a refinement network. In the design of the antenna array, we add an additional design level and employ a hypernetwork within a hypernetwork. The learning objective is based on measuring the similarity of the obtained radiation pattern to the desired one. Our experiments demonstrate that our approach is able to design novel antennas and antenna arrays that are compliant with the design requirements, considerably better than the baseline methods. We compare the solutions obtained by our method to existing designs and demonstrate a high level of overlap. When designing the antenna array of a cellular phone, the obtained solution displays improved properties over the existing one.
****
Value Iteration in Continuous Actions, States and Time
Michael Lutter, Shie Mannor, Jan Peters, Dieter Fox, Animesh Garg
Classical value iteration approaches are not applicable to environments with continuous states and actions. For such environments the states and actions must be discretized, which leads to an exponential increase in computational complexity. In this paper, we propose continuous fitted value iteration (cFVI). This algorithm enables dynamic programming for continuous states and actions with a known dynamics model. Exploiting the continuous time formulation, the optimal policy can be derived for non-linear control-affine dynamics. This closed-form solution enables the efficient extension of value iteration to continuous environments. We show in non-linear control experiments that the dynamic programming solution obtains the same quantitative performance as deep reinforcement learning methods in simulation but excels when transferred to the physical system.The policy obtained by cFVI is more robust to changes in the dynamics despite using only a deterministic model and without explicitly incorporating robustness in the optimization
****
Meta-Cal: Well-controlled Post-hoc Calibration by Ranking
Xingchen Ma, Matthew B. Blaschko
In many applications, it is desirable that a classifier not only makes accurate predictions, but also outputs calibrated posterior probabilities. However, many existing classifiers, especially deep neural network classifiers, tend to be uncalibrated. Post-hoc calibration is a technique to recalibrate a model by learning a calibration map. Existing approaches mostly focus on constructing calibration maps with low calibration errors, however, this quality is inadequate for a calibrator being useful. In this paper, we introduce two constraints that are worth consideration in designing a calibration map for post-hoc calibration. Then we present Meta-Cal, which is built from a base calibrator and a ranking model. Under some mild assumptions, two high-probability bounds are given with respect to these constraints. Empirical results on CIFAR-10, CIFAR-100 and ImageNet and a range of popular network architectures show our proposed method significantly outperforms the current state of the art for post-hoc multi-class classification calibration.
****
Neural-Pull: Learning Signed Distance Function from Point clouds by Learning to Pull Space onto Surface
Baorui Ma, Zhizhong Han, Yu-Shen Liu, Matthias Zwicker
Reconstructing continuous surfaces from 3D point clouds is a fundamental operation in 3D geometry processing. Several recent state-of-the-art methods address this problem using neural networks to learn signed distance functions (SDFs). In this paper, we introduce Neural-Pull, a new approach that is simple and leads to high quality SDFs. Specifically, we train a neural network to pull query 3D locations to their closest points on the surface using the predicted signed distance values and the gradient at the query locations, both of which are computed by the network itself. The pulling operation moves each query location with a stride given by the distance predicted by the network. Based on the sign of the distan

ce, this may move the query location along or against the direction of the gradient of the SDF. This is a differentiable operation that allows us to update the signed distance value and the gradient simultaneously during training. Our outperforming results under widely used benchmarks demonstrate that we can learn SDFs more accurately and flexibly for surface reconstruction and single image reconstruction than the state-of-the-art methods. Our code and data are available at https://github.com/mabaorui/NeuralPull.
****
Learning Stochastic Behaviour from Aggregate Data
Shaojun Ma, Shu Liu, Hongyuan Zha, Haomin Zhou
Learning nonlinear dynamics from aggregate data is a challenging problem because the full trajectory of each individual is not available, namely, the individual observed at one time may not be observed at the next time point, or the identity of individual is unavailable. This is in sharp contrast to learning dynamics with full trajectory data, on which the majority of existing methods are based. We propose a novel method using the weak form of Fokker Planck Equation (FPE) — a partial differential equation — to describe the density evolution of data in a sampled form, which is then combined with Wasserstein generative adversarial network (WGAN) in the training process. In such a sample-based framework we are able to learn the nonlinear dynamics from aggregate data without explicitly solving the partial differential equation (PDE) FPE. We demonstrate our approach in the context of a series of synthetic and real-world data sets.
****
Local Algorithms for Finding Densely Connected Clusters
Peter Macgregor, He Sun
Local graph clustering is an important algorithmic technique for analysing massive graphs, and has been widely applied in many research fields of data science. While the objective of most (local) graph clustering algorithms is to find a vertex set of low conductance, there has been a sequence of recent studies that highlight the importance of the inter-connection between clusters when analysing real-world datasets. Following this line of research, in this work we study local algorithms for finding a pair of vertex sets defined with respect to their inter-connection and their relationship with the rest of the graph. The key to our analysis is a new reduction technique that relates the structure of multiple sets to a single vertex set in the reduced graph. Among many potential applications, we show that our algorithms successfully recover densely connected clusters in the Interstate Disputes Dataset and the US Migration Dataset.
****
Learning to Generate Noise for Multi-Attack Robustness
Divyam Madaan, Jinwoo Shin, Sung Ju Hwang
Adversarial learning has emerged as one of the successful techniques to circumvent the susceptibility of existing methods against adversarial perturbations. However, the majority of existing defense methods are tailored to defend against a single category of adversarial perturbation (e.g. ℓ_∞-attack). In safety-critical applications, this makes these methods extraneous as the attacker can adopt diverse adversaries to deceive the system. Moreover, training on multiple perturbations simultaneously significantly increases the computational overhead during training. To address these challenges, we propose a novel meta-learning framework that explicitly learns to generate noise to improve the model's robustness against multiple types of attacks. Its key component is \emph{Meta Noise Generator (MNG)} that outputs optimal noise to stochastically perturb a given sample, such that it helps lower the error on diverse adversarial perturbations. By utilizing samples generated by MNG, we train a model by enforcing the label consistency across multiple perturbations. We validate the robustness of models trained by our scheme on various datasets and against a wide variety of perturbations, demonstrating that it significantly outperforms the baselines across multiple perturbations with a marginal computational cost.
****
Learning Interaction Kernels for Agent Systems on Riemannian Manifolds
Mauro Maggioni, Jason J Miller, Hongda Qiu, Ming Zhong

Interacting agent and particle systems are extensively used to model complex phe nomena in science and engineering. We consider the problem of learning interacti on kernels in these dynamical systems constrained to evolve on Riemannian manifo lds from given trajectory data. The models we consider are based on interaction kernels depending on pairwise Riemannian distances between agents, with agents i nteracting locally along the direction of the shortest geodesic connecting them. We show that our estimators converge at a rate that is independent of the dimen sion of the state space, and derive bounds on the trajectory estimation error, o n the manifold, between the observed and estimated dynamics. We demonstrate the performance of our estimator on two classical first order interacting systems: O pinion Dynamics and a Predator-Swarm system, with each system constrained on two prototypical manifolds, the 2-dimensional sphere and the Poincaré disk model of hyperbolic space.
****
Tesseract: Tensorised Actors for Multi-Agent Reinforcement Learning
Anuj Mahajan, Mikayel Samvelyan, Lei Mao, Viktor Makoviychuk, Animesh Garg, Jean Kossaifi, Shimon Whiteson, Yuke Zhu, Animashree Anandkumar
Reinforcement Learning in large action spaces is a challenging problem. This is especially true for cooperative multi-agent reinforcement learning (MARL), which often requires tractable learning while respecting various constraints like com munication budget and information about other agents. In this work, we focus on the fundamental hurdle affecting both value-based and policy-gradient approaches : an exponential blowup of the action space with the number of agents. For value -based methods, it poses challenges in accurately representing the optimal value function for value-based methods, thus inducing suboptimality. For policy gradi ent methods, it renders the critic ineffective and exacerbates the problem of th e lagging critic. We show that from a learning theory perspective, both problems can be addressed by accurately representing the associated action-value functio n with a low-complexity hypothesis class. This requires accurately modelling the agent interactions in a sample efficient way. To this end, we propose a novel t ensorised formulation of the Bellman equation. This gives rise to our method Tes seract, which utilises the view of Q-function seen as a tensor where the modes c orrespond to action spaces of different agents. Algorithms derived from Tesserac t decompose the Q-tensor across the agents and utilise low-rank tensor approxima tions to model the agent interactions relevant to the task. We provide PAC analy sis for Tesseract based algorithms and highlight their relevance to the class of rich observation MDPs. Empirical results in different domains confirm the gains in sample efficiency using Tesseract as supported by the theory.
****
Domain Generalization using Causal Matching
Divyat Mahajan, Shruti Tople, Amit Sharma
In the domain generalization literature, a common objective is to learn represen tations independent of the domain after conditioning on the class label. We show that this objective is not sufficient: there exist counter-examples where a mod el fails to generalize to unseen domains even after satisfying class-conditional domain invariance. We formalize this observation through a structural causal mo del and show the importance of modeling within-class variations for generalizati on. Specifically, classes contain objects that characterize specific causal feat ures, and domains can be interpreted as interventions on these objects that chan ge non-causal features. We highlight an alternative condition: inputs across dom ains should have the same representation if they are derived from the same objec t. Based on this objective, we propose matching-based algorithms when base objec ts are observed (e.g., through data augmentation) and approximate the objective when objects are not observed (MatchDG). Our simple matching-based algorithms ar e competitive to prior work on out-of-domain accuracy for rotated MNIST, Fashion -MNIST, PACS, and Chest-Xray datasets. Our method MatchDG also recovers ground-t ruth object matches: on MNIST and Fashion-MNIST, top-10 matches from MatchDG hav e over 50% overlap with ground-truth matches.
****
Stability and Convergence of Stochastic Gradient Clipping: Beyond Lipschitz Cont

inuity and Smoothness
Vien V. Mai, Mikael Johansson
Stochastic gradient algorithms are often unstable when applied to functions that do not have Lipschitz-continuous and/or bounded gradients. Gradient clipping is a simple and effective technique to stabilize the training process for problems that are prone to the exploding gradient problem. Despite its widespread popularity, the convergence properties of the gradient clipping heuristic are poorly understood, especially for stochastic problems. This paper establishes both qualitative and quantitative convergence results of the clipped stochastic (sub)gradient method (SGD) for non-smooth convex functions with rapidly growing subgradients. Our analyses show that clipping enhances the stability of SGD and that the clipped SGD algorithm enjoys finite convergence rates in many cases. We also study the convergence of a clipped method with momentum, which includes clipped SGD as a special case, for weakly convex problems under standard assumptions. With a novel Lyapunov analysis, we show that the proposed method achieves the best-known rate for the considered class of problems, demonstrating the effectiveness of clipped methods also in this regime. Numerical results confirm our theoretical developments.
****
Nonparametric Hamiltonian Monte Carlo
Carol Mak, Fabian Zaiser, Luke Ong
Probabilistic programming uses programs to express generative models whose posterior probability is then computed by built-in inference engines. A challenging goal is to develop general purpose inference algorithms that work out-of-the-box for arbitrary programs in a universal probabilistic programming language (PPL). The densities defined by such programs, which may use stochastic branching and recursion, are (in general) nonparametric, in the sense that they correspond to models on an infinite-dimensional parameter space. However standard inference algorithms, such as the Hamiltonian Monte Carlo (HMC) algorithm, target distributions with a fixed number of parameters. This paper introduces the Nonparametric Hamiltonian Monte Carlo (NP-HMC) algorithm which generalises HMC to nonparametric models. Inputs to NP-HMC are a new class of measurable functions called “tree representable”, which serve as a language-independent representation of the density functions of probabilistic programs in a universal PPL. We provide a correctness proof of NP-HMC, and empirically demonstrate significant performance improvements over existing approaches on several nonparametric examples.
****
Exploiting structured data for learning contagious diseases under incomplete testing
Maggie Makar, Lauren West, David Hooper, Eric Horvitz, Erica Shenoy, John Guttag
One of the ways that machine learning algorithms can help control the spread of an infectious disease is by building models that predict who is likely to become infected making them good candidates for preemptive interventions. In this work we ask: can we build reliable infection prediction models when the observed data is collected under limited, and biased testing that prioritizes testing symptomatic individuals? Our analysis suggests that when the infection is highly transmissible, incomplete testing might be sufficient to achieve good out-of-sample prediction error. Guided by this insight, we develop an algorithm that predicts infections, and show that it outperforms baselines on simulated data. We apply our model to data from a large hospital to predict Clostridioides difficile infections; a communicable disease that is characterized by both symptomatically infected and asymptomatic (i.e., untested) carriers. Using a proxy instead of the unobserved untested-infected state, we show that our model outperforms benchmarks in predicting infections.
****
Near-Optimal Algorithms for Explainable k-Medians and k-Means
Konstantin Makarychev, Liren Shan
We consider the problem of explainable k-medians and k-means introduced by Dasgupta, Frost, Moshkovitz, and Rashtchian (ICML 2020). In this problem, our goal is to find a \emph{threshold decision tree} that partitions data into k clus

ters and minimizes the k-medians or k-means objective. The obtained clustering is easy to interpret because every decision node of a threshold tree splits data based on a single feature into two groups. We propose a new algorithm for this problem which is $\tilde O(\log k)$ competitive with k-medians with ℓ_1 norm and $\tilde O(k)$ competitive with k-means. This is an improvement over the previous guarantees of $O(k)$ and $O(k^2)$ by Dasgupta et al (2020). We also provide a new algorithm which is $O(\log^{\nicefrac{3}{2}} k)$ competitive for k-medians with ℓ_2 norm. Our first algorithm is near-optimal: Dasgupta et al (2020) showed a lower bound of $\Omega(\log k)$ for k-medians; in this work, we prove a lower bound of $\tilde\Omega(k)$ for k-means. We also provide a lower bound of $\Omega(\log k)$ for k-medians with ℓ_2 norm.
****
KO codes: inventing nonlinear encoding and decoding for reliable wireless communication via deep-learning
Ashok V Makkuva, Xiyang Liu, Mohammad Vahid Jamali, Hessam Mahdavifar, Sewoong Oh, Pramod Viswanath
Landmark codes underpin reliable physical layer communication, e.g., Reed-Muller, BCH, Convolution, Turbo, LDPC, and Polar codes: each is a linear code and represents a mathematical breakthrough. The impact on humanity is huge: each of these codes has been used in global wireless communication standards (satellite, WiFi, cellular). Reliability of communication over the classical additive white Gaussian noise (AWGN) channel enables benchmarking and ranking of the different codes. In this paper, we construct KO codes, a computationally efficient family of deep-learning driven (encoder, decoder) pairs that outperform the state-of-the-art reliability performance on the standardized AWGN channel. KO codes beat state-of-the-art Reed-Muller and Polar codes, under the low-complexity successive cancellation decoding, in the challenging short-to-medium block length regime on the AWGN channel. We show that the gains of KO codes are primarily due to the nonlinear mapping of information bits directly to transmit symbols (bypassing modulation) and yet possess an efficient, high-performance decoder. The key technical innovation that renders this possible is design of a novel family of neural architectures inspired by the computation tree of the {\bf K}ronecker {\bf O}peration (KO) central to Reed-Muller and Polar codes. These architectures pave way for the discovery of a much richer class of hitherto unexplored nonlinear algebraic structures.
****
Quantifying the Benefit of Using Differentiable Learning over Tangent Kernels
Eran Malach, Pritish Kamath, Emmanuel Abbe, Nathan Srebro
We study the relative power of learning with gradient descent on differentiable models, such as neural networks, versus using the corresponding tangent kernels. We show that under certain conditions, gradient descent achieves small error only if a related tangent kernel method achieves a non-trivial advantage over random guessing (a.k.a. weak learning), though this advantage might be very small even when gradient descent can achieve arbitrarily high accuracy. Complementing this, we show that without these conditions, gradient descent can in fact learn with small error even when no kernel method, in particular using the tangent kernel, can achieve a non-trivial advantage over random guessing.
****
Inverse Constrained Reinforcement Learning
Shehryar Malik, Usman Anwar, Alireza Aghasi, Ali Ahmed
In real world settings, numerous constraints are present which are hard to specify mathematically. However, for the real world deployment of reinforcement learning (RL), it is critical that RL agents are aware of these constraints, so that they can act safely. In this work, we consider the problem of learning constraints from demonstrations of a constraint-abiding agent’s behavior. We experimentally validate our approach and show that our framework can successfully learn the most likely constraints that the agent respects. We further show that these learned constraints are \textit{transferable} to new agents that may have different morphologies and/or reward functions. Previous works in this regard have either mainly been restricted to tabular (discrete) settings, specific types of constra

ints or assume the environment’s transition dynamics. In contrast, our framework is able to learn arbitrary \textit{Markovian} constraints in high-dimensions in a completely model-free setting. The code is available at: \url{https://github.com/shehryar-malik/icrl}.
****
A Sampling-Based Method for Tensor Ring Decomposition
Osman Asif Malik, Stephen Becker
We propose a sampling-based method for computing the tensor ring (TR) decomposition of a data tensor. The method uses leverage score sampled alternating least squares to fit the TR cores in an iterative fashion. By taking advantage of the special structure of TR tensors, we can efficiently estimate the leverage scores and attain a method which has complexity sublinear in the number of input tensor entries. We provide high-probability relative-error guarantees for the sampled least squares problems. We compare our proposal to existing methods in experiments on both synthetic and real data. Our method achieves substantial speedup—sometimes two or three orders of magnitude—over competing methods, while maintaining good accuracy. We also provide an example of how our method can be used for rapid feature extraction.
****
Sample Efficient Reinforcement Learning In Continuous State Spaces: A Perspective Beyond Linearity
Dhruv Malik, Aldo Pacchiano, Vishwak Srinivasan, Yuanzhi Li
Reinforcement learning (RL) is empirically successful in complex nonlinear Markov decision processes (MDPs) with continuous state spaces. By contrast, the majority of theoretical RL literature requires the MDP to satisfy some form of linear structure, in order to guarantee sample efficient RL. Such efforts typically assume the transition dynamics or value function of the MDP are described by linear functions of the state features. To resolve this discrepancy between theory and practice, we introduce the Effective Planning Window (EPW) condition, a structural condition on MDPs that makes no linearity assumptions. We demonstrate that the EPW condition permits sample efficient RL, by providing an algorithm which provably solves MDPs satisfying this condition. Our algorithm requires minimal assumptions on the policy class, which can include multi-layer neural networks with nonlinear activation functions. Notably, the EPW condition is directly motivated by popular gaming benchmarks, and we show that many classic Atari games satisfy this condition. We additionally show the necessity of conditions like EPW, by demonstrating that simple MDPs with slight nonlinearities cannot be solved sample efficiently.
****
Beyond the Pareto Efficient Frontier: Constraint Active Search for Multiobjective Experimental Design
Gustavo Malkomes, Bolong Cheng, Eric H Lee, Mike Mccourt
Many problems in engineering design and simulation require balancing competing objectives under the presence of uncertainty. Sample-efficient multiobjective optimization methods focus on the objective function values in metric space and ignore the sampling behavior of the design configurations in parameter space. Consequently, they may provide little actionable insight on how to choose designs in the presence of metric uncertainty or limited precision when implementing a chosen design. We propose a new formulation that accounts for the importance of the parameter space and is thus more suitable for multiobjective design problems; instead of searching for the Pareto-efficient frontier, we solicit the desired minimum performance thresholds on all objectives to define regions of satisfaction. We introduce an active search algorithm called Expected Coverage Improvement (ECI) to efficiently discover the region of satisfaction and simultaneously sample diverse acceptable configurations. We demonstrate our algorithm on several design and simulation domains: mechanical design, additive manufacturing, medical monitoring, and plasma physics.
****
Consistent Nonparametric Methods for Network Assisted Covariate Estimation
Xueyu Mao, Deepayan Chakrabarti, Purnamrita Sarkar

Networks with node covariates are commonplace: for example, people in a social network have interests, or product preferences, etc. If we know the covariates for some nodes, can we infer them for the remaining nodes? In this paper we propose a new similarity measure between two nodes based on the patterns of their 2-hop neighborhoods. We show that a simple algorithm (CN-VEC) like nearest neighbor regression with this metric is consistent for a wide range of models when the degree grows faster than $n^{1/3}$ up-to logarithmic factors, where n is the number of nodes. For "low-rank" latent variable models, the natural contender will be to estimate the latent variables using SVD and use them for non-parametric regression. While we show consistency of this method under less stringent sparsity conditions, our experimental results suggest that the simple local CN-VEC method either outperforms the global SVD-RBF method, or has comparable performance for low rank models. We also present simulated and real data experiments to show the effectiveness of our algorithms compared to the state of the art.
****
Near-Optimal Model-Free Reinforcement Learning in Non-Stationary Episodic MDPs
Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, Tamer Basar
We consider model-free reinforcement learning (RL) in non-stationary Markov decision processes. Both the reward functions and the state transition functions are allowed to vary arbitrarily over time as long as their cumulative variations do not exceed certain variation budgets. We propose Restarted Q-Learning with Upper Confidence Bounds (RestartQ-UCB), the first model-free algorithm for non-stationary RL, and show that it outperforms existing solutions in terms of dynamic regret. Specifically, RestartQ-UCB with Freedman-type bonus terms achieves a dynamic regret bound of $\widetilde{O}(S^{\frac{1}{3}} A^{\frac{1}{3}} \Delta^{\frac{1}{3}} H T^{\frac{2}{3}})$, where S and A are the numbers of states and actions, respectively, $\Delta>0$ is the variation budget, H is the number of time steps per episode, and T is the total number of time steps. We further show that our algorithm is \emph{nearly optimal} by establishing an information-theoretical lower bound of $\Omega(S^{\frac{1}{3}} A^{\frac{1}{3}} \Delta^{\frac{1}{3}} H^{\frac{2}{3}} T^{\frac{2}{3}})$, the first lower bound in non-stationary RL. Numerical experiments validate the advantages of RestartQ-UCB in terms of both cumulative rewards and computational efficiency. We further demonstrate the power of our results in the context of multi-agent RL, where non-stationarity is a key challenge.
****
Adaptive Sampling for Best Policy Identification in Markov Decision Processes
Aymen Al Marjani, Alexandre Proutiere
We investigate the problem of best-policy identification in discounted Markov Decision Processes (MDPs) when the learner has access to a generative model. The objective is to devise a learning algorithm returning the best policy as early as possible. We first derive a problem-specific lower bound of the sample complexity satisfied by any learning algorithm. This lower bound corresponds to an optimal sample allocation that solves a non-convex program, and hence, is hard to exploit in the design of efficient algorithms. We then provide a simple and tight upper bound of the sample complexity lower bound, whose corresponding nearly-optimal sample allocation becomes explicit. The upper bound depends on specific functionals of the MDP such as the sub-optimality gaps and the variance of the next-state value function, and thus really captures the hardness of the MDP. Finally, we devise KLB-TS (KL Ball Track-and-Stop), an algorithm tracking this nearly-optimal allocation, and provide asymptotic guarantees for its sample complexity (both almost surely and in expectation). The advantages of KLB-TS against state-of-the-art algorithms are discussed and illustrated numerically.
****
Explanations for Monotonic Classifiers.
Joao Marques-Silva, Thomas Gerspacher, Martin C Cooper, Alexey Ignatiev, Nina Narodytska
In many classification tasks there is a requirement of monotonicity. Concretely, if all else remains constant, increasing (resp. decreasing) the value of one or more features must not decrease (resp. increase) the value of the prediction. D

espite comprehensive efforts on learning monotonic classifiers, dedicated approaches for explaining monotonic classifiers are scarce and classifier-specific. This paper describes novel algorithms for the computation of one formal explanation of a (black-box) monotonic classifier. These novel algorithms are polynomial (indeed linear) in the run time complexity of the classifier. Furthermore, the paper presents a practically efficient model-agnostic algorithm for enumerating formal explanations.
****
Multi-Agent Training beyond Zero-Sum with Correlated Equilibrium Meta-Solvers
Luke Marris, Paul Muller, Marc Lanctot, Karl Tuyls, Thore Graepel
Two-player, constant-sum games are well studied in the literature, but there has been limited progress outside of this setting. We propose Joint Policy-Space Response Oracles (JPSRO), an algorithm for training agents in n-player, general-sum extensive form games, which provably converges to an equilibrium. We further suggest correlated equilibria (CE) as promising meta-solvers, and propose a novel solution concept Maximum Gini Correlated Equilibrium (MGCE), a principled and computationally efficient family of solutions for solving the correlated equilibrium selection problem. We conduct several experiments using CE meta-solvers for JPSRO and demonstrate convergence on n-player, general-sum games.
****
Blind Pareto Fairness and Subgroup Robustness
Natalia L Martinez, Martin A Bertran, Afroditi Papadaki, Miguel Rodrigues, Guillermo Sapiro
Much of the work in the field of group fairness addresses disparities between predefined groups based on protected features such as gender, age, and race, which need to be available at train, and often also at test, time. These approaches are static and retrospective, since algorithms designed to protect groups identified a priori cannot anticipate and protect the needs of different at-risk groups in the future. In this work we analyze the space of solutions for worst-case fairness beyond demographics, and propose Blind Pareto Fairness (BPF), a method that leverages no-regret dynamics to recover a fair minimax classifier that reduces worst-case risk of any potential subgroup of sufficient size, and guarantees that the remaining population receives the best possible level of service. BPF addresses fairness beyond demographics, that is, it does not rely on predefined notions of at-risk groups, neither at train nor at test time. Our experimental results show that the proposed framework improves worst-case risk in multiple standard datasets, while simultaneously providing better levels of service for the remaining population. The code is available at github.com/natalialmg/BlindParetoFairness
****
Necessary and sufficient conditions for causal feature selection in time series with latent common causes
Atalanti A Mastakouri, Bernhard Schölkopf, Dominik Janzing
We study the identification of direct and indirect causes on time series with latent variables, and provide a constrained-based causal feature selection method, which we prove that is both sound and complete under some graph constraints. Our theory and estimation algorithm require only two conditional independence tests for each observed candidate time series to determine whether or not it is a cause of an observed target time series. Furthermore, our selection of the conditioning set is such that it improves signal to noise ratio. We apply our method on real data, and on a wide range of simulated experiments, which yield very low false positive and relatively low false negative rates.
****
Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restriction
Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, Krikamol Muandet
We address the problem of causal effect estima-tion in the presence of unobserved confounding,but where proxies for the latent confounder(s) areobserved. We propose two kernel-based meth-ods for nonlinear causal effect estimation in thisset

ting: (a) a two-stage regression approach, and(b) a maximum moment restriction a pproach. Wefocus on the proximal causal learning setting, butour methods can be used to solve a wider classof inverse problems characterised by a Fredholmintegr al equation. In particular, we provide a uni-fying view of two-stage and moment restrictionapproaches for solving this problem in a nonlin-ear setting. We provi de consistency guaranteesfor each algorithm, and demonstrate that these ap-proac hes achieve competitive results on syntheticdata and data simulating a real-worl d task. In par-ticular, our approach outperforms earlier methodsthat are not sui ted to leveraging proxy variables.
****
Robust Unsupervised Learning via L-statistic Minimization
Andreas Maurer, Daniela Angela Parletta, Andrea Paudice, Massimiliano Pontil
Designing learning algorithms that are resistant to perturbations of the underly ing data distribution is a problem of wide practical and theoretical importance. We present a general approach to this problem focusing on unsupervised learning . The key assumption is that the perturbing distribution is characterized by lar ger losses relative to a given class of admissible models. This is exploited by a general descent algorithm which minimizes an L-statistic criterion over the model class, weighting small losses more. Our analysis characterizes the robustn ess of the method in terms of bounds on the reconstruction error relative to the underlying unperturbed distribution. As a byproduct, we prove uniform convergen ce bounds with respect to the proposed criterion for several popular models in u nsupervised learning, a result which may be of independent interest. Numerical e xperiments with \textsc{kmeans} clustering and principal subspace analysis demon strate the effectiveness of our approach.
****
Adversarial Multi Class Learning under Weak Supervision with Performance Guarant ees
Alessio Mazzetto, Cyrus Cousins, Dylan Sam, Stephen H Bach, Eli Upfal
We develop a rigorous approach for using a set of arbitrarily correlated weak su pervision sources in order to solve a multiclass classification task when only a very small set of labeled data is available. Our learning algorithm provably co nverges to a model that has minimum empirical risk with respect to an adversaria l choice over feasible labelings for a set of unlabeled data, where the feasibil ity of a labeling is computed through constraints defined by rigorously estimate d statistics of the weak supervision sources. We show theoretical guarantees for this approach that depend on the information provided by the weak supervision s ources. Notably, this method does not require the weak supervision sources to ha ve the same labeling space as the multiclass classification task. We demonstrate the effectiveness of our approach with experiments on various image classificat ion tasks.
****
Fundamental Tradeoffs in Distributionally Adversarial Training
Mohammad Mehrabi, Adel Javanmard, Ryan A. Rossi, Anup Rao, Tung Mai
Adversarial training is among the most effective techniques to improve robustnes s of models against adversarial perturbations. However, the full effect of this approach on models is not well understood. For example, while adversarial traini ng can reduce the adversarial risk (prediction error against an adversary), it s ometimes increase standard risk (generalization error when there is no adversary). In this paper, we focus on \emph{distribution perturbing} adversary framework wherein the adversary can change the test distribution within a neighborhood of the training data distribution. The neighborhood is defined via Wasserstein dis tance between distributions and the radius of the neighborhood is a measure of a dversary’s manipulative power. We study the tradeoff between standard risk and a dversarial risk and derive the Pareto-optimal tradeoff, achievable over specific classes of models, in the infinite data limit with features dimension kept fixe d. We consider three learning settings: 1) Regression with the class of linear m odels; 2) Binary classification under the Gaussian mixtures data model, with the class of linear classifiers; 3) Regression with the class of random features mo del (which can be equivalently represented as two-layer neural network with rand

om first-layer weights). We show that a tradeoff between standard and adversaria
l risk is manifested in all three settings. We further characterize the Pareto-o
ptimal tradeoff curves and discuss how a variety of factors, such as features co
rrelation, adversary’s power or the width of two-layer neural network would affe
ct this tradeoff.
****
Leveraging Non-uniformity in First-order Non-convex Optimization
Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, Dale Schuurmans
Classical global convergence results for first-order methods rely on uniform smo
othness and the ■{}ojasiewicz inequality. Motivated by properties of objective f
unctions that arise in machine learning, we propose a non-uniform refinement of
these notions, leading to \emph{Non-uniform Smoothness} (NS) and \emph{Non-unifo
rm ■{}ojasiewicz inequality} (N■{}). The new definitions inspire new geometry-aw
are first-order methods that are able to converge to global optimality faster th
an the classical $\Omega(1/t^2)$ lower bounds. To illustrate the power of these
geometry-aware methods and their corresponding non-uniform analysis, we consider
 two important problems in machine learning: policy gradient optimization in rei
nforcement learning (PG), and generalized linear model training in supervised le
arning (GLM). For PG, we find that normalizing the gradient ascent method can ac
celerate convergence to $O(e^{- c \cdot t})$ (where $c > 0$) while incurring les
s overhead than existing algorithms. For GLM, we show that geometry-aware normal
ized gradient descent can also achieve a linear convergence rate, which signific
antly improves the best known results. We additionally show that the proposed ge
ometry-aware gradient descent methods escape landscape plateaus faster than stan
dard gradient descent. Experimental results are used to illustrate and complemen
t the theoretical findings.
****
Controlling Graph Dynamics with Reinforcement Learning and Graph Neural Networks
Eli Meirom, Haggai Maron, Shie Mannor, Gal Chechik
We consider the problem of controlling a partially-observed dynamic process on a
 graph by a limited number of interventions. This problem naturally arises in co
ntexts such as scheduling virus tests to curb an epidemic; targeted marketing in
 order to promote a product; and manually inspecting posts to detect fake news s
preading on social networks. We formulate this setup as a sequential decision pr
oblem over a temporal graph process. In face of an exponential state space, comb
inatorial action space and partial observability, we design a novel tractable sc
heme to control dynamical processes on temporal graphs. We successfully apply ou
r approach to two popular problems that fall into our framework: prioritizing wh
ich nodes should be tested in order to curb the spread of an epidemic, and influ
ence maximization on a graph.
****
A theory of high dimensional regression with arbitrary correlations between inpu
t features and target functions: sample complexity, multiple descent curves and
a hierarchy of phase transitions
Gabriel Mel, Surya Ganguli
The performance of neural networks depends on precise relationships between four
 distinct ingredients: the architecture, the loss function, the statistical stru
cture of inputs, and the ground truth target function. Much theoretical work has
 focused on understanding the role of the first two ingredients under highly sim
plified models of random uncorrelated data and target functions. In contrast, pe
rformance likely relies on a conspiracy between the statistical structure of the
 input distribution and the structure of the function to be learned. To understa
nd this better we revisit ridge regression in high dimensions, which corresponds
 to an exceedingly simple architecture and loss function, but we analyze its per
formance under arbitrary correlations between input features and the target func
tion. We find a rich mathematical structure that includes: (1) a dramatic reduct
ion in sample complexity when the target function aligns with data anisotropy; (
2) the existence of multiple descent curves; (3) a sequence of phase transitions
 in the performance, loss landscape, and optimal regularization as a function of
 the amount of data that explains the first two effects.

****
Neural Architecture Search without Training
Joe Mellor, Jack Turner, Amos Storkey, Elliot J Crowley
The time and effort involved in hand-designing deep neural networks is immense. This has prompted the development of Neural Architecture Search (NAS) techniques to automate this design. However, NAS algorithms tend to be slow and expensive; they need to train vast numbers of candidate networks to inform the search process. This could be alleviated if we could partially predict a network’s trained accuracy from its initial state. In this work, we examine the overlap of activations between datapoints in untrained networks and motivate how this can give a measure which is usefully indicative of a network’s trained performance. We incorporate this measure into a simple algorithm that allows us to search for powerful networks without any training in a matter of seconds on a single GPU, and verify its effectiveness on NAS-Bench-101, NAS-Bench-201, NATS-Bench, and Network Design Spaces. Our approach can be readily combined with more expensive search methods; we examine a simple adaptation of regularised evolutionary search. Code for reproducing our experiments is available at https://github.com/BayesWatch/nas-without-training.
****
Fast active learning for pure exploration in reinforcement learning
Pierre Menard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, Michal Valko
Realistic environments often provide agents with very limited feedback. When the environment is initially unknown, the feedback, in the beginning, can be completely absent, and the agents may first choose to devote all their effort on \emph{exploring efficiently.} The exploration remains a challenge while it has been addressed with many hand-tuned heuristics with different levels of generality on one side, and a few theoretically-backed exploration strategies on the other. Many of them are incarnated by \emph{intrinsic motivation} and in particular \emph{explorations bonuses}. A common choice is to use $1/\sqrt{n}$ bonus, where n is a number of times this particular state-action pair was visited. We show that, surprisingly, for a pure-exploration objective of \emph{reward-free exploration}, bonuses that scale with $1/n$ bring faster learning rates, improving the known upper bounds with respect to the dependence on the horizon H. Furthermore, we show that with an improved analysis of the stopping time, we can improve by a factor H the sample complexity in the \emph{best-policy identification} setting, which is another pure-exploration objective, where the environment provides rewards but the agent is not penalized for its behavior during the exploration phase.
****
UCB Momentum Q-learning: Correcting the bias without forgetting
Pierre Menard, Omar Darwiche Domingues, Xuedong Shang, Michal Valko
We propose UCBMQ, Upper Confidence Bound Momentum Q-learning, a new algorithm for reinforcement learning in tabular and possibly stage-dependent, episodic Markov decision process. UCBMQ is based on Q-learning where we add a momentum term and rely on the principle of optimism in face of uncertainty to deal with exploration. Our new technical ingredient of UCBMQ is the use of momentum to correct the bias that Q-learning suffers while, \emph{at the same time}, limiting the impact it has on the second-order term of the regret. For UCBMQ, we are able to guarantee a regret of at most $\tilde{O}(\sqrt{H^3SAT}+ H^4 S A)$ where H is the length of an episode, S the number of states, A the number of actions, T the number of episodes and ignoring terms in poly$\log(SAHT)$. Notably, UCBMQ is the first algorithm that simultaneously matches the lower bound of $\Omega(\sqrt{H^3SAT})$ for large enough T and has a second-order term (with respect to T) that scales \emph{only linearly} with the number of states S.
****
An Integer Linear Programming Framework for Mining Constraints from Data
Tao Meng, Kai-Wei Chang
Structured output prediction problems (e.g., sequential tagging, hierarchical multi-class classification) often involve constraints over the output space. These

constraints interact with the learned models to filter infeasible solutions and facilitate in building an accountable system. However, despite constraints are useful, they are often based on hand-crafted rules. This raises a question – can we mine constraints and rules from data based on a learning algorithm? In this paper, we present a general framework for mining constraints from data. In particular, we consider the inference in structured output prediction as an integer linear programming (ILP) problem. Then, given the coefficients of the objective function and the corresponding solution, we mine the underlying constraints by estimating the outer and inner polytopes of the feasible set. We verify the proposed constraint mining algorithm in various synthetic and real-world applications and demonstrate that the proposed approach successfully identifies the feasible set at scale. In particular, we show that our approach can learn to solve 9x9 Sudoku puzzles and minimal spanning tree problems from examples without providing the underlying rules. Our algorithm can also integrate with a neural network model to learn the hierarchical label structure of a multi-label classification task. Besides, we provide theoretical analysis about the tightness of the polytopes and the reliability of the mined constraints.
****

A statistical perspective on distillation

Aditya K Menon, Ankit Singh Rawat, Sashank Reddi, Seungyeon Kim, Sanjiv Kumar

Knowledge distillation is a technique for improving a “student” model by replacing its one-hot training labels with a label distribution obtained from a “teacher” model. Despite its broad success, several basic questions — e.g., Why does distillation help? Why do more accurate teachers not necessarily distill better? — have received limited formal study. In this paper, we present a statistical perspective on distillation which provides an answer to these questions. Our core observation is that a “Bayes teacher” providing the true class-probabilities can lower the variance of the student objective, and thus improve performance. We then establish a bias-variance tradeoff that quantifies the value of teachers that approximate the Bayes class-probabilities. This provides a formal criterion as to what constitutes a “good” teacher, namely, the quality of its probability estimates. Finally, we illustrate how our statistical perspective facilitates novel applications of distillation to bipartite ranking and multiclass retrieval.
****

Learn2Hop: Learned Optimization on Rough Landscapes

Amil Merchant, Luke Metz, Samuel S Schoenholz, Ekin D Cubuk

Optimization of non-convex loss surfaces containing many local minima remains a critical problem in a variety of domains, including operations research, informatics, and material design. Yet, current techniques either require extremely high iteration counts or a large number of random restarts for good performance. In this work, we propose adapting recent developments in meta-learning to these many-minima problems by learning the optimization algorithm for various loss landscapes. We focus on problems from atomic structural optimization—finding low energy configurations of many-atom systems—including widely studied models such as bimetallic clusters and disordered silicon. We find that our optimizer learns a hopping behavior which enables efficient exploration and improves the rate of low energy minima discovery. Finally, our learned optimizers show promising generalization with efficiency gains on never before seen tasks (e.g. new elements or compositions). Code is available at https://learn2hop.page.link/github.
****

Counterfactual Credit Assignment in Model-Free Reinforcement Learning

Thomas Mesnard, Theophane Weber, Fabio Viola, Shantanu Thakoor, Alaa Saade, Anna Harutyunyan, Will Dabney, Thomas S Stepleton, Nicolas Heess, Arthur Guez, Eric Moulines, Marcus Hutter, Lars Buesing, Remi Munos

Credit assignment in reinforcement learning is the problem of measuring an action’s influence on future rewards. In particular, this requires separating skill from luck, i.e. disentangling the effect of an action on rewards from that of external factors and subsequent actions. To achieve this, we adapt the notion of counterfactuals from causality theory to a model-free RL setup. The key idea is to condition value functions on future events, by learning to extract relevant inf

ormation from a trajectory. We formulate a family of policy gradient algorithms that use these future-conditional value functions as baselines or critics, and show that they are provably low variance. To avoid the potential bias from conditioning on future information, we constrain the hindsight information to not contain information about the agent's actions. We demonstrate the efficacy and validity of our algorithm on a number of illustrative and challenging problems.
****

Provably Efficient Learning of Transferable Rewards
Alberto Maria Metelli, Giorgia Ramponi, Alessandro Concetti, Marcello Restelli
The reward function is widely accepted as a succinct, robust, and transferable representation of a task. Typical approaches, at the basis of Inverse Reinforcement Learning (IRL), leverage on expert demonstrations to recover a reward function. In this paper, we study the theoretical properties of the class of reward functions that are compatible with the expert's behavior. We analyze how the limited knowledge of the expert's policy and of the environment affects the reward reconstruction phase. Then, we examine how the error propagates to the learned policy's performance when transferring the reward function to a different environment. We employ these findings to devise a provably efficient active sampling approach, aware of the need for transferring the reward function, that can be paired with a large variety of IRL algorithms. Finally, we provide numerical simulations on benchmark environments.
****

Mixed Nash Equilibria in the Adversarial Examples Game
Laurent Meunier, Meyer Scetbon, Rafael B Pinot, Jamal Atif, Yann Chevaleyre
This paper tackles the problem of adversarial examples from a game theoretic point of view. We study the open question of the existence of mixed Nash equilibria in the zero-sum game formed by the attacker and the classifier. While previous works usually allow only one player to use randomized strategies, we show the necessity of considering randomization for both the classifier and the attacker. We demonstrate that this game has no duality gap, meaning that it always admits approximate Nash equilibria. We also provide the first optimization algorithms to learn a mixture of classifiers that approximately realizes the value of this game, \emph{i.e.} procedures to build an optimally robust randomized classifier.
****

Learning in Nonzero-Sum Stochastic Games with Potentials
David H Mguni, Yutong Wu, Yali Du, Yaodong Yang, Ziyi Wang, Minne Li, Ying Wen, Joel Jennings, Jun Wang
Multi-agent reinforcement learning (MARL) has become effective in tackling discrete cooperative game scenarios. However, MARL has yet to penetrate settings beyond those modelled by team and zero-sum games, confining it to a small subset of multi-agent systems. In this paper, we introduce a new generation of MARL learners that can handle \textit{nonzero-sum} payoff structures and continuous settings. In particular, we study the MARL problem in a class of games known as stochastic potential games (SPGs) with continuous state-action spaces. Unlike cooperative games, in which all agents share a common reward, SPGs are capable of modelling real-world scenarios where agents seek to fulfil their individual goals. We prove theoretically our learning method, $\ourmethod$, enables independent agents to learn Nash equilibrium strategies in \textit{polynomial time}. We demonstrate our framework tackles previously unsolvable tasks such as \textit{Coordination Navigation} and \textit{large selfish routing games} and that it outperforms the state of the art MARL baselines such as MADDPG and COMIX in such scenarios.
****

EfficientTTS: An Efficient and High-Quality Text-to-Speech Architecture
Chenfeng Miao, Liang Shuang, Zhengchen Liu, Chen Minchuan, Jun Ma, Shaojun Wang, Jing Xiao
In this work, we address the Text-to-Speech (TTS) task by proposing a non-autoregressive architecture called EfficientTTS. Unlike the dominant non-autoregressive TTS models, which are trained with the need of external aligners, EfficientTTS optimizes all its parameters with a stable, end-to-end training procedure, allowing for synthesizing high quality speech in a fast and efficient manner. Effici

entTTS is motivated by a new monotonic alignment modeling approach, which specif
ies monotonic constraints to the sequence alignment with almost no increase of c
omputation. By combining EfficientTTS with different feed-forward network struct
ures, we develop a family of TTS models, including both text-to-melspectrogram a
nd text-to-waveform networks. We experimentally show that the proposed models si
gnificantly outperform counterpart models such as Tacotron 2 and Glow-TTS in ter
ms of speech quality, training efficiency and synthesis speed, while still produ
cing the speeches of strong robustness and great diversity. In addition, we demo
nstrate that proposed approach can be easily extended to autoregressive models s
uch as Tacotron 2.
****
Outside the Echo Chamber: Optimizing the Performative Risk
John P Miller, Juan C Perdomo, Tijana Zrnic
In performative prediction, predictions guide decision-making and hence can infl
uence the distribution of future data. To date, work on performative prediction
has focused on finding performatively stable models, which are the fixed points
of repeated retraining. However, stable solutions can be far from optimal when e
valuated in terms of the performative risk, the loss experienced by the decision
 maker when deploying a model. In this paper, we shift attention beyond performa
tive stability and focus on optimizing the performative risk directly. We identi
fy a natural set of properties of the loss function and model-induced distributi
on shift under which the performative risk is convex, a property which does not
follow from convexity of the loss alone. Furthermore, we develop algorithms that
 leverage our structural assumptions to optimize the performative risk with bett
er sample efficiency than generic methods for derivative-free convex optimizatio
n.
****
Accuracy on the Line: on the Strong Correlation Between Out-of-Distribution and
In-Distribution Generalization
John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vais
haal Shankar, Percy Liang, Yair Carmon, Ludwig Schmidt
For machine learning systems to be reliable, we must understand their performanc
e in unseen, out- of-distribution environments. In this paper, we empirically sh
ow that out-of-distribution performance is strongly correlated with in-distribut
ion performance for a wide range of models and distribution shifts. Specifically
, we demonstrate strong correlations between in-distribution and out-of- distrib
ution performance on variants of CIFAR- 10 & ImageNet, a synthetic pose estimati
on task derived from YCB objects, FMoW-WILDS satellite imagery classification, a
nd wildlife classification in iWildCam-WILDS. The correlation holds across model
 architectures, hyperparameters, training set size, and training duration, and i
s more precise than what is expected from existing domain adaptation theory. To
complete the picture, we also investigate cases where the correlation is weaker,
 for instance some synthetic distribution shifts from CIFAR-10-C and the tissue
classification dataset Camelyon17-WILDS. Finally, we provide a candidate theory
based on a Gaussian data model that shows how changes in the data covariance ari
sing from distribution shift can affect the observed correlations.
****
Signatured Deep Fictitious Play for Mean Field Games with Common Noise
Ming Min, Ruimeng Hu
Existing deep learning methods for solving mean-field games (MFGs) with common n
oise fix the sampling common noise paths and then solve the corresponding MFGs.
This leads to a nested loop structure with millions of simulations of common noi
se paths in order to produce accurate solutions, which results in prohibitive co
mputational cost and limits the applications to a large extent. In this paper, b
ased on the rough path theory, we propose a novel single-loop algorithm, named s
ignatured deep fictitious play (Sig-DFP), by which we can work with the unfixed
common noise setup to avoid the nested loop structure and reduce the computation
al complexity significantly. The proposed algorithm can accurately capture the e
ffect of common uncertainty changes on mean-field equilibria without further tra
ining of neural networks, as previously needed in the existing machine learning

algorithms. The efficiency is supported by three applications, including linear-quadratic MFGs, mean-field portfolio game, and mean-field game of optimal consumption and investment. Overall, we provide a new point of view from the rough path theory to solve MFGs with common noise with significantly improved efficiency and an extensive range of applications. In addition, we report the first deep learning work to deal with extended MFGs (a mean-field interaction via both the states and controls) with common noise.
****
Meta-StyleSpeech : Multi-Speaker Adaptive Text-to-Speech Generation
Dongchan Min, Dong Bok Lee, Eunho Yang, Sung Ju Hwang
With rapid progress in neural text-to-speech (TTS) models, personalized speech generation is now in high demand for many applications. For practical applicability, a TTS model should generate high-quality speech with only a few audio samples from the given speaker, that are also short in length. However, existing methods either require to fine-tune the model or achieve low adaptation quality without fine-tuning. In this work, we propose StyleSpeech, a new TTS model which not only synthesizes high-quality speech but also effectively adapts to new speakers. Specifically, we propose Style-Adaptive Layer Normalization (SALN) which aligns gain and bias of the text input according to the style extracted from a reference speech audio. With SALN, our model effectively synthesizes speech in the style of the target speaker even from a single speech audio. Furthermore, to enhance StyleSpeech’s adaptation to speech from new speakers, we extend it to Meta-StyleSpeech by introducing two discriminators trained with style prototypes, and performing episodic training. The experimental results show that our models generate high-quality speech which accurately follows the speaker’s voice with single short-duration (1-3 sec) speech audio, significantly outperforming baselines.
****
On the Explicit Role of Initialization on the Convergence and Implicit Bias of Overparametrized Linear Networks
Hancheng Min, Salma Tarmoun, Rene Vidal, Enrique Mallada
Neural networks trained via gradient descent with random initialization and without any regularization enjoy good generalization performance in practice despite being highly overparametrized. A promising direction to explain this phenomenon is to study how initialization and overparametrization affect convergence and implicit bias of training algorithms. In this paper, we present a novel analysis of single-hidden-layer linear networks trained under gradient flow, which connects initialization, optimization, and overparametrization. Firstly, we show that the squared loss converges exponentially to its optimum at a rate that depends on the level of imbalance of the initialization. Secondly, we show that proper initialization constrains the dynamics of the network parameters to lie within an invariant set. In turn, minimizing the loss over this set leads to the min-norm solution. Finally, we show that large hidden layer width, together with (properly scaled) random initialization, ensures proximity to such an invariant set during training, allowing us to derive a novel non-asymptotic upper-bound on the distance between the trained network and the min-norm solution.
****
An Identifiable Double VAE For Disentangled Representations
Graziano Mita, Maurizio Filippone, Pietro Michiardi
A large part of the literature on learning disentangled representations focuses on variational autoencoders (VAEs). Recent developments demonstrate that disentanglement cannot be obtained in a fully unsupervised setting without inductive biases on models and data. However, Khemakhem et al., AISTATS, 2020 suggest that employing a particular form of factorized prior, conditionally dependent on auxiliary variables complementing input observations, can be one such bias, resulting in an identifiable model with guarantees on disentanglement. Working along this line, we propose a novel VAE-based generative model with theoretical guarantees on identifiability. We obtain our conditional prior over the latents by learning an optimal representation, which imposes an additional strength on their regularization. We also extend our method to semi-supervised settings. Experimental results indicate superior performance with respect to state-of-the-art approaches

, according to several established metrics proposed in the literature on disentanglement.
****
Offline Meta-Reinforcement Learning with Advantage Weighting
Eric Mitchell, Rafael Rafailov, Xue Bin Peng, Sergey Levine, Chelsea Finn
This paper introduces the offline meta-reinforcement learning (offline meta-RL) problem setting and proposes an algorithm that performs well in this setting. Offline meta-RL is analogous to the widely successful supervised learning strategy of pre-training a model on a large batch of fixed, pre-collected data (possibly from various tasks) and fine-tuning the model to a new task with relatively little data. That is, in offline meta-RL, we meta-train on fixed, pre-collected data from several tasks and adapt to a new task with a very small amount (less than 5 trajectories) of data from the new task. By nature of being offline, algorithms for offline meta-RL can utilize the largest possible pool of training data available and eliminate potentially unsafe or costly data collection during meta-training. This setting inherits the challenges of offline RL, but it differs significantly because offline RL does not generally consider a) transfer to new tasks or b) limited data from the test task, both of which we face in offline meta-RL. Targeting the offline meta-RL setting, we propose Meta-Actor Critic with Advantage Weighting (MACAW). MACAW is an optimization-based meta-learning algorithm that uses simple, supervised regression objectives for both the inner and outer loop of meta-training. On offline variants of common meta-RL benchmarks, we empirically find that this approach enables fully offline meta-reinforcement learning and achieves notable gains over prior methods.
****
The Power of Log-Sum-Exp: Sequential Density Ratio Matrix Estimation for Speed-Accuracy Optimization
Taiki Miyagawa, Akinori F Ebihara
We propose a model for multiclass classification of time series to make a prediction as early and as accurate as possible. The matrix sequential probability ratio test (MSPRT) is known to be asymptotically optimal for this setting, but contains a critical assumption that hinders broad real-world applications; the MSPRT requires the underlying probability density. To address this problem, we propose to solve density ratio matrix estimation (DRME), a novel type of density ratio estimation that consists of estimating matrices of multiple density ratios with constraints and thus is more challenging than the conventional density ratio estimation. We propose a log-sum-exp-type loss function (LSEL) for solving DRME and prove the following: (i) the LSEL provides the true density ratio matrix as the sample size of the training set increases (consistency); (ii) it assigns larger gradients to harder classes (hard class weighting effect); and (iii) it provides discriminative scores even on class-imbalanced datasets (guess-aversion). Our overall architecture for early classification, MSPRT-TANDEM, statistically significantly outperforms baseline models on four datasets including action recognition, especially in the early stage of sequential observations. Our code and datasets are publicly available.
****
PODS: Policy Optimization via Differentiable Simulation
Miguel Angel Zamora Mora, Momchil Peychev, Sehoon Ha, Martin Vechev, Stelian Coros
Current reinforcement learning (RL) methods use simulation models as simple black-box oracles. In this paper, with the goal of improving the performance exhibited by RL algorithms, we explore a systematic way of leveraging the additional information provided by an emerging class of differentiable simulators. Building on concepts established by Deterministic Policy Gradients (DPG) methods, the neural network policies learned with our approach represent deterministic actions. In a departure from standard methodologies, however, learning these policies does not hinge on approximations of the value function that must be learned concurrently in an actor-critic fashion. Instead, we exploit differentiable simulators to directly compute the analytic gradient of a policy’s value function with respect to the actions it outputs. This, in turn, allows us to efficiently perform lo

cally optimal policy improvement iterations. Compared against other state-of-the-art RL methods, we show that with minimal hyper-parameter tuning our approach consistently leads to better asymptotic behavior across a set of payload manipulation tasks that demand a high degree of accuracy and precision.
****
Efficient Deviation Types and Learning for Hindsight Rationality in Extensive-Form Games
Dustin Morrill, Ryan D'Orazio, Marc Lanctot, James R Wright, Michael Bowling, Amy R Greenwald
Hindsight rationality is an approach to playing general-sum games that prescribes no-regret learning dynamics for individual agents with respect to a set of deviations, and further describes jointly rational behavior among multiple agents with mediated equilibria. To develop hindsight rational learning in sequential decision-making settings, we formalize behavioral deviations as a general class of deviations that respect the structure of extensive-form games. Integrating the idea of time selection into counterfactual regret minimization (CFR), we introduce the extensive-form regret minimization (EFR) algorithm that achieves hindsight rationality for any given set of behavioral deviations with computation that scales closely with the complexity of the set. We identify behavioral deviation subsets, the partial sequence deviation types, that subsume previously studied types and lead to efficient EFR instances in games with moderate lengths. In addition, we present a thorough empirical analysis of EFR instantiated with different deviation types in benchmark games, where we find that stronger types typically induce better performance.
****
Neural Rough Differential Equations for Long Time Series
James Morrill, Cristopher Salvi, Patrick Kidger, James Foster
Neural controlled differential equations (CDEs) are the continuous-time analogue of recurrent neural networks, as Neural ODEs are to residual networks, and offer a memory-efficient continuous-time way to model functions of potentially irregular time series. Existing methods for computing the forward pass of a Neural CDE involve embedding the incoming time series into path space, often via interpolation, and using evaluations of this path to drive the hidden state. Here, we use rough path theory to extend this formulation. Instead of directly embedding into path space, we instead represent the input signal over small time intervals through its \textit{log-signature}, which are statistics describing how the signal drives a CDE. This is the approach for solving \textit{rough differential equations} (RDEs), and correspondingly we describe our main contribution as the introduction of Neural RDEs. This extension has a purpose: by generalising the Neural CDE approach to a broader class of driving signals, we demonstrate particular advantages for tackling long time series. In this regime, we demonstrate efficacy on problems of length up to 17k observations and observe significant training speed-ups, improvements in model performance, and reduced memory requirements compared to existing approaches.
****
Connecting Interpretability and Robustness in Decision Trees through Separation
Michal Moshkovitz, Yao-Yuan Yang, Kamalika Chaudhuri
Recent research has recognized interpretability and robustness as essential properties of trustworthy classification. Curiously, a connection between robustness and interpretability was empirically observed, but the theoretical reasoning behind it remained elusive. In this paper, we rigorously investigate this connection. Specifically, we focus on interpretation using decision trees and robustness to l_{\infty}-perturbation. Previous works defined the notion of r-separation as a sufficient condition for robustness. We prove upper and lower bounds on the tree size in case the data is r-separated. We then show that a tighter bound on the size is possible when the data is linearly separated. We provide the first algorithm with provable guarantees both on robustness, interpretability, and accuracy in the context of decision trees. Experiments confirm that our algorithm yields classifiers that are both interpretable and robust and have high accuracy.
****

Outlier-Robust Optimal Transport
Debarghya Mukherjee, Aritra Guha, Justin M Solomon, Yuekai Sun, Mikhail Yurochki
n
Optimal transport (OT) measures distances between distributions in a way that de
pends on the geometry of the sample space. In light of recent advances in comput
ational OT, OT distances are widely used as loss functions in machine learning.
Despite their prevalence and advantages, OT loss functions can be extremely sens
itive to outliers. In fact, a single adversarially-picked outlier can increase t
he standard W_2-distance arbitrarily. To address this issue, we propose an out
lier-robust formulation of OT. Our formulation is convex but challenging to scal
e at a first glance. Our main contribution is deriving an \emph{equivalent} form
ulation based on cost truncation that is easy to incorporate into modern algorit
hms for computational OT. We demonstrate the benefits of our formulation in mean
 estimation problems under the Huber contamination model in simulations and outl
ier detection tasks on real data.
****
Oblivious Sketching for Logistic Regression
Alexander Munteanu, Simon Omlor, David Woodruff
What guarantees are possible for solving logistic regression in one pass over a
data stream? To answer this question, we present the first data oblivious sketch
 for logistic regression. Our sketch can be computed in input sparsity time over
 a turnstile data stream and reduces the size of a d-dimensional data set from
 n to only $\operatorname{poly}(\mu d\log n)$ weighted points, where μ is
a useful parameter which captures the complexity of compressing the data. Solvin
g (weighted) logistic regression on the sketch gives an $O(\log n)$-approximatio
n to the original problem on the full data set. We also show how to obtain an $O
(1)$-approximation with slight modifications. Our sketches are fast, simple, eas
y to implement, and our experiments demonstrate their practicality.
****
Bias-Variance Reduced Local SGD for Less Heterogeneous Federated Learning
Tomoya Murata, Taiji Suzuki
Recently, local SGD has got much attention and been extensively studied in the d
istributed learning community to overcome the communication bottleneck problem.
However, the superiority of local SGD to minibatch SGD only holds in quite limit
ed situations. In this paper, we study a new local algorithm called Bias-Varianc
e Reduced Local SGD (BVR-L-SGD) for nonconvex distributed optimization. Algorith
mically, our proposed bias and variance reduced local gradient estimator fully u
tilizes small second-order heterogeneity of local objectives and suggests random
ly picking up one of the local models instead of taking the average of them when
 workers are synchronized. Theoretically, under small heterogeneity of local obj
ectives, we show that BVR-L-SGD achieves better communication complexity than bo
th the previous non-local and local methods under mild conditions, and particula
rly BVR-L-SGD is the first method that breaks the barrier of communication compl
exity $\Theta(1/\varepsilon)$ for general nonconvex smooth objectives when the h
eterogeneity is small and the local computation budget is large. Numerical resul
ts are given to verify the theoretical findings and give empirical evidence of t
he superiority of our method.
****
Implicit-PDF: Non-Parametric Representation of Probability Distributions on the
Rotation Manifold
Kieran A Murphy, Carlos Esteves, Varun Jampani, Srikumar Ramalingam, Ameesh Maka
dia
In the deep learning era, the vast majority of methods to predict pose from a si
ngle image are trained to classify or regress to a single given ground truth pos
e per image. Such methods have two main shortcomings, i) they cannot represent u
ncertainty about the predictions, and ii) they cannot handle symmetric objects,
where multiple (potentially infinite) poses may be correct. Only recently these
shortcomings have been addressed, but current approaches as limited in that they
 cannot express the full rich space of distributions on the rotation manifold. T
o this end, we introduce a method to estimate arbitrary, non-parametric distribu

tions on SO(3). Our key idea is to represent the distributions implicitly, with a neural network that estimates the probability density, given the input image and a candidate pose. At inference time, grid sampling or gradient ascent can be used to find the most likely pose, but it is also possible to evaluate the density at any pose, enabling reasoning about symmetries and uncertainty. This is the most general way of representing distributions on manifolds, and to demonstrate its expressive power we introduce a new dataset containing symmetric and nearly-symmetric objects. Our method also shows advantages on the popular object pose estimation benchmarks ModelNet10-SO(3) and T-LESS. Code, data, and visualizations may be found at implicit-pdf.github.io.
****

No-regret Algorithms for Capturing Events in Poisson Point Processes
Mojmir Mutny, Andreas Krause
Inhomogeneous Poisson point processes are widely used models of event occurrences. We address \emph{adaptive sensing of Poisson Point processes}, namely, maximizing the number of captured events subject to sensing costs. We encode prior assumptions on the rate function by modeling it as a member of a known \emph{reproducing kernel Hilbert space} (RKHS). By partitioning the domain into separate small regions, and using heteroscedastic linear regression, we propose a tractable estimator of Poisson process rates for two feedback models: \emph{count-record}, where exact locations of events are observed, and \emph{histogram} feedback, where only counts of events are observed. We derive provably accurate anytime confidence estimates for our estimators for sequentially acquired Poisson count data. Using these, we formulate algorithms based on optimism that provably incur sublinear count-regret. We demonstrate the practicality of the method on problems from crime modeling, revenue maximization as well as environmental monitoring.
****

Online Limited Memory Neural-Linear Bandits with Likelihood Matching
Ofir Nabati, Tom Zahavy, Shie Mannor
We study neural-linear bandits for solving problems where {\em both} exploration and representation learning play an important role. Neural-linear bandits harnesses the representation power of Deep Neural Networks (DNNs) and combines it with efficient exploration mechanisms by leveraging uncertainty estimation of the model, designed for linear contextual bandits on top of the last hidden layer. In order to mitigate the problem of representation change during the process, new uncertainty estimations are computed using stored data from an unlimited buffer. Nevertheless, when the amount of stored data is limited, a phenomenon called catastrophic forgetting emerges. To alleviate this, we propose a likelihood matching algorithm that is resilient to catastrophic forgetting and is completely online. We applied our algorithm, Limited Memory Neural-Linear with Likelihood Matching (NeuralLinear-LiM2) on a variety of datasets and observed that our algorithm achieves comparable performance to the unlimited memory approach while exhibits resilience to catastrophic forgetting.
****

Quantitative Understanding of VAE as a Non-linearly Scaled Isometric Embedding
Akira Nakagawa, Keizo Kato, Taiji Suzuki
Variational autoencoder (VAE) estimates the posterior parameters (mean and variance) of latent variables corresponding to each input data. While it is used for many tasks, the transparency of the model is still an underlying issue. This paper provides a quantitative understanding of VAE property through the differential geometric and information-theoretic interpretations of VAE. According to the Rate-distortion theory, the optimal transform coding is achieved by using an orthonormal transform with PCA basis where the transform space is isometric to the input. Considering the analogy of transform coding to VAE, we clarify theoretically and experimentally that VAE can be mapped to an implicit isometric embedding with a scale factor derived from the posterior parameter. As a result, we can estimate the data probabilities in the input space from the prior, loss metrics, and corresponding posterior parameters, and further, the quantitative importance of each latent variable can be evaluated like the eigenvalue of PCA.
****

GMAC: A Distributional Perspective on Actor-Critic Framework
Daniel W Nam, Younghoon Kim, Chan Y Park
In this paper, we devise a distributional framework on actor-critic as a solutio
n to distributional instability, action type restriction, and conflation between
 samples and statistics. We propose a new method that minimizes the Cram{é}r dis
tance with the multi-step Bellman target distribution generated from a novel Sam
ple-Replacement algorithm denoted SR(\lambda), which learns the correct value di
stribution under multiple Bellman operations. Parameterizing a value distributio
n with Gaussian Mixture Model further improves the efficiency and the performanc
e of the method, which we name GMAC. We empirically show that GMAC captures the
correct representation of value distributions and improves the performance of a
conventional actor-critic method with low computational cost, in both discrete a
nd continuous action spaces using Arcade Learning Environment (ALE) and PyBullet
 environment.
****
Memory-Efficient Pipeline-Parallel DNN Training
Deepak Narayanan, Amar Phanishayee, Kaiyu Shi, Xie Chen, Matei Zaharia
Many state-of-the-art ML results have been obtained by scaling up the number of
parameters in existing models. However, parameters and activations for such larg
e models often do not fit in the memory of a single accelerator device; this mea
ns that it is necessary to distribute training of large models over multiple acc
elerators. In this work, we propose PipeDream-2BW, a system that supports memory
-efficient pipeline parallelism. PipeDream-2BW uses a novel pipelining and weigh
t gradient coalescing strategy, combined with the double buffering of weights, t
o ensure high throughput, low memory footprint, and weight update semantics simi
lar to data parallelism. In addition, PipeDream-2BW automatically partitions the
 model over the available hardware resources, while respecting hardware constrai
nts such as memory capacities of accelerators and interconnect topologies. PipeD
ream-2BW can accelerate the training of large GPT and BERT language models by up
 to 20x with similar final model accuracy.
****
Randomized Dimensionality Reduction for Facility Location and Single-Linkage Clu
stering
Shyam Narayanan, Sandeep Silwal, Piotr Indyk, Or Zamir
Random dimensionality reduction is a versatile tool for speeding up algorithms f
or high-dimensional problems. We study its application to two clustering problem
s: the facility location problem, and the single-linkage hierarchical clustering
 problem, which is equivalent to computing the minimum spanning tree. We show th
at if we project the input pointset X onto a random $d = O(d_X)$-dimensional s
ubspace (where d_X is the doubling dimension of X), then the optimum facilit
y location cost in the projected space approximates the original cost up to a co
nstant factor. We show an analogous statement for minimum spanning tree, but wit
h the dimension d having an extra $\log \log n$ term and the approximation fac
tor being arbitrarily close to 1. Furthermore, we extend these results to appr
oximating {\em solutions} instead of just their {\em costs}. Lastly, we provide
experimental results to validate the quality of solutions and the speedup due to
 the dimensionality reduction. Unlike several previous papers studying this appr
oach in the context of k-means and k-medians, our dimension bound does not d
epend on the number of clusters but only on the intrinsic dimensionality of X.
****
Generating images with sparse representations
Charlie Nash, Jacob Menick, Sander Dieleman, Peter Battaglia
The high dimensionality of images presents architecture and sampling-efficiency
challenges for likelihood-based generative models. Previous approaches such as V
Q-VAE use deep autoencoders to obtain compact representations, which are more pr
actical as inputs for likelihood-based models. We present an alternative approac
h, inspired by common image compression methods like JPEG, and convert images to
 quantized discrete cosine transform (DCT) blocks, which are represented sparsel
y as a sequence of DCT channel, spatial location, and DCT coefficient triples. W
e propose a Transformer-based autoregressive architecture, which is trained to s

equentially predict the conditional distribution of the next element in such sequences, and which scales effectively to high resolution images. On a range of image datasets, we demonstrate that our approach can generate high quality, diverse images, with sample metric scores competitive with state of the art methods. We additionally show that simple modifications to our method yield effective image colorization and super-resolution models.
****
Geometric convergence of elliptical slice sampling
Viacheslav Natarovskii, Daniel Rudolf, Björn Sprungk
For Bayesian learning, given likelihood function and Gaussian prior, the elliptical slice sampler, introduced by Murray, Adams and MacKay 2010, provides a tool for the construction of a Markov chain for approximate sampling of the underlying posterior distribution. Besides of its wide applicability and simplicity its main feature is that no tuning is necessary. Under weak regularity assumptions on the posterior density we show that the corresponding Markov chain is geometrically ergodic and therefore yield qualitative convergence guarantees. We illustrate our result for Gaussian posteriors as they appear in Gaussian process regression in a fully Gaussian scenario, which for example is exhibited in Gaussian process regression, as well as in a setting of a multi-modal distribution. Remarkably, our numerical experiments indicate a dimension-independent performance of elliptical slice sampling even in situations where our ergodicity result does not apply.
****
HardCoRe-NAS: Hard Constrained diffeRentiable Neural Architecture Search
Niv Nayman, Yonathan Aflalo, Asaf Noy, Lihi Zelnik
Realistic use of neural networks often requires adhering to multiple constraints on latency, energy and memory among others. A popular approach to find fitting networks is through constrained Neural Architecture Search (NAS), however, previous methods enforce the constraint only softly. Therefore, the resulting networks do not exactly adhere to the resource constraint and their accuracy is harmed. In this work we resolve this by introducing Hard Constrained diffeRentiable NAS (HardCoRe-NAS), that is based on an accurate formulation of the expected resource requirement and a scalable search method that satisfies the hard constraint throughout the search. Our experiments show that HardCoRe-NAS generates state-of-the-art architectures, surpassing other NAS methods, while strictly satisfying the hard resource constraints without any tuning required.
****
Emergent Social Learning via Multi-agent Reinforcement Learning
Kamal K Ndousse, Douglas Eck, Sergey Levine, Natasha Jaques
Social learning is a key component of human and animal intelligence. By taking cues from the behavior of experts in their environment, social learners can acquire sophisticated behavior and rapidly adapt to new circumstances. This paper investigates whether independent reinforcement learning (RL) agents in a multi-agent environment can learn to use social learning to improve their performance. We find that in most circumstances, vanilla model-free RL agents do not use social learning. We analyze the reasons for this deficiency, and show that by imposing constraints on the training environment and introducing a model-based auxiliary loss we are able to obtain generalized social learning policies which enable agents to: i) discover complex skills that are not learned from single-agent training, and ii) adapt online to novel environments by taking cues from experts present in the new environment. In contrast, agents trained with model-free RL or imitation learning generalize poorly and do not succeed in the transfer tasks. By mixing multi-agent and solo training, we can obtain agents that use social learning to gain skills that they can deploy when alone, even out-performing agents trained alone from the start.
****
Bayesian Algorithm Execution: Estimating Computable Properties of Black-box Functions Using Mutual Information
Willie Neiswanger, Ke Alexander Wang, Stefano Ermon
In many real world problems, we want to infer some property of an expensive blac

k-box function f, given a budget of T function evaluations. One example is budge t constrained global optimization of f, for which Bayesian optimization is a pop ular method. Other properties of interest include local optima, level sets, inte grals, or graph-structured information induced by f. Often, we can find an algor ithm A to compute the desired property, but it may require far more than T queri es to execute. Given such an A, and a prior distribution over f, we refer to the problem of inferring the output of A using T evaluations as Bayesian Algorithm Execution (BAX). To tackle this problem, we present a procedure, InfoBAX, that s equentially chooses queries that maximize mutual information with respect to the algorithm's output. Applying this to Dijkstra's algorithm, for instance, we inf er shortest paths in synthetic and real-world graphs with black-box edge costs. Using evolution strategies, we yield variants of Bayesian optimization that targ et local, rather than global, optima. On these problems, InfoBAX uses up to 500 times fewer queries to f than required by the original algorithm. Our method is closely connected to other Bayesian optimal experimental design procedures such as entropy search methods and optimal sensor placement using Gaussian processes.
****
Continuous Coordination As a Realistic Scenario for Lifelong Learning
Hadi Nekoei, Akilesh Badrinaaraayanan, Aaron Courville, Sarath Chandar
Current deep reinforcement learning (RL) algorithms are still highly task-specif ic and lack the ability to generalize to new environments. Lifelong learning (LL L), however, aims at solving multiple tasks sequentially by efficiently transfer ring and using knowledge between tasks. Despite a surge of interest in lifelong RL in recent years, the lack of a realistic testbed makes robust evaluation of L LL algorithms difficult. Multi-agent RL (MARL), on the other hand, can be seen a s a natural scenario for lifelong RL due to its inherent non-stationarity, since the agents' policies change over time. In this work, we introduce a multi-agent lifelong learning testbed that supports both zero-shot and few-shot settings. O ur setup is based on Hanabi {—} a partially-observable, fully cooperative multi- agent game that has been shown to be challenging for zero-shot coordination. Its large strategy space makes it a desirable environment for lifelong RL tasks. We evaluate several recent MARL methods, and benchmark state-of-the-art LLL algori thms in limited memory and computation regimes to shed light on their strengths and weaknesses. This continual learning paradigm also provides us with a pragmat ic way of going beyond centralized training which is the most commonly used trai ning protocol in MARL. We empirically show that the agents trained in our setup are able to coordinate well with unseen agents, without any additional assumptio ns made by previous works. The code and all pre-trained models are available at https://github.com/chandar-lab/Lifelong-Hanabi.
****
Policy Caches with Successor Features
Mark Nemecek, Ronald Parr
Transfer in reinforcement learning is based on the idea that it is possible to u se what is learned in one task to improve the learning process in another task. For transfer between tasks which share transition dynamics but differ in reward function, successor features have been shown to be a useful representation which allows for efficient computation of action-value functions for previously-learn ed policies in new tasks. These functions induce policies in the new tasks, so a n agent may not need to learn a new policy for each new task it encounters, espe cially if it is allowed some amount of suboptimality in those tasks. We present new bounds for the performance of optimal policies in a new task, as well as an approach to use these bounds to decide, when presented with a new task, whether to use cached policies or learn a new policy.
****
Causality-aware counterfactual confounding adjustment as an alternative to linea r residualization in anticausal prediction tasks based on linear learners
Elias Chaibub Neto
Linear residualization is a common practice for confounding adjustment in machin e learning applications. Recently, causality-aware predictive modeling has been proposed as an alternative causality-inspired approach for adjusting for confoun

ders. In this paper, we compare the linear residualization approach against the causality-aware confounding adjustment in anticausal prediction tasks. Our comparisons include both the settings where the training and test sets come from the same distributions, as well as, when the training and test sets are shifted due to selection biases. In the absence of dataset shifts, we show that the causality-aware approach tends to (asymptotically) outperform the residualization adjustment in terms of predictive performance in linear learners. Importantly, our results still holds even when the true model generating the data is not linear. We illustrate our results in both regression and classification tasks. Furthermore, in the presence of dataset shifts in the joint distribution of the confounders and outcome variables, we show that the causality-aware approach is more stable than linear residualization.
****
Incentivizing Compliance with Algorithmic Instruments
Dung Daniel T Ngo, Logan Stapleton, Vasilis Syrgkanis, Steven Wu
Randomized experiments can be susceptible to selection bias due to potential non-compliance by the participants. While much of the existing work has studied compliance as a static behavior, we propose a game-theoretic model to study compliance as dynamic behavior that may change over time. In rounds, a social planner interacts with a sequence of heterogeneous agents who arrive with their unobserved private type that determines both their prior preferences across the actions (e.g., control and treatment) and their baseline rewards without taking any treatment. The planner provides each agent with a randomized recommendation that may alter their beliefs and their action selection. We develop a novel recommendation mechanism that views the planner’s recommendation as a form of instrumental variable (IV) that only affects an agents’ action selection, but not the observed rewards. We construct such IVs by carefully mapping the history –the interactions between the planner and the previous agents– to a random recommendation. Even though the initial agents may be completely non-compliant, our mechanism can incentivize compliance over time, thereby enabling the estimation of the treatment effect of each treatment, and minimizing the cumulative regret of the planner whose goal is to identify the optimal treatment.
****
On the Proof of Global Convergence of Gradient Descent for Deep ReLU Networks with Linear Widths
Quynh Nguyen
We give a simple proof for the global convergence of gradient descent in training deep ReLU networks with the standard square loss, and show some of its improvements over the state-of-the-art. In particular, while prior works require all the hidden layers to be wide with width at least $\Omega(N^8)$ (N being the number of training samples), we require a single wide layer of linear, quadratic or cubic width depending on the type of initialization. Unlike many recent proofs based on the Neural Tangent Kernel (NTK), our proof need not track the evolution of the entire NTK matrix, or more generally, any quantities related to the changes of activation patterns during training. Instead, we only need to track the evolution of the output at the last hidden layer, which can be done much more easily thanks to the Lipschitz property of ReLU. Some highlights of our setting: (i) all the layers are trained with standard gradient descent, (ii) the network has standard parameterization as opposed to the NTK one, and (iii) the network has a single wide layer as opposed to having all wide hidden layers as in most of NTK-related results.
****
Value-at-Risk Optimization with Gaussian Processes
Quoc Phong Nguyen, Zhongxiang Dai, Bryan Kian Hsiang Low, Patrick Jaillet
Value-at-risk (VaR) is an established measure to assess risks in critical real-world applications with random environmental factors. This paper presents a novel VaR upper confidence bound (V-UCB) algorithm for maximizing the VaR of a black-box objective function with the first no-regret guarantee. To realize this, we first derive a confidence bound of VaR and then prove the existence of values of the environmental random variable (to be selected to achieve no regret) such tha

t the confidence bound of VaR lies within that of the objective function evaluated at such values. Our V-UCB algorithm empirically demonstrates state-of-the-art performance in optimizing synthetic benchmark functions, a portfolio optimization problem, and a simulated robot task.
****
Cross-model Back-translated Distillation for Unsupervised Machine Translation
Xuan-Phi Nguyen, Shafiq Joty, Thanh-Tung Nguyen, Kui Wu, Ai Ti Aw
Recent unsupervised machine translation (UMT) systems usually employ three main principles: initialization, language modeling and iterative back-translation, though they may apply them differently. Crucially, iterative back-translation and denoising auto-encoding for language modeling provide data diversity to train the UMT systems. However, the gains from these diversification processes has seemed to plateau. We introduce a novel component to the standard UMT framework called Cross-model Back-translated Distillation (CBD), that is aimed to induce another level of data diversification that existing principles lack. CBD is applicable to all previous UMT approaches. In our experiments, CBD achieves the state of the art in the WMT’14 English-French, WMT’16 English-German and English-Romanian bilingual unsupervised translation tasks, with 38.2, 30.1, and 36.3 BLEU respectively. It also yields 1.5–3.3 BLEU improvements in IWSLT English-French and English-German tasks. Through extensive experimental analyses, we show that CBD is effective because it embraces data diversity while other similar variants do not.
****
Optimal Transport Kernels for Sequential and Parallel Neural Architecture Search
Vu Nguyen, Tam Le, Makoto Yamada, Michael A. Osborne
Neural architecture search (NAS) automates the design of deep neural networks. One of the main challenges in searching complex and non-continuous architectures is to compare the similarity of networks that the conventional Euclidean metric may fail to capture. Optimal transport (OT) is resilient to such complex structure by considering the minimal cost for transporting a network into another. However, the OT is generally not negative definite which may limit its ability to build the positive-definite kernels required in many kernel-dependent frameworks. Building upon tree-Wasserstein (TW), which is a negative definite variant of OT, we develop a novel discrepancy for neural architectures, and demonstrate it within a Gaussian process surrogate model for the sequential NAS settings. Furthermore, we derive a novel parallel NAS, using quality k-determinantal point process on the GP posterior, to select diverse and high-performing architectures from a discrete set of candidates. Empirically, we demonstrate that our TW-based approaches outperform other baselines in both sequential and parallel NAS.
****
Interactive Learning from Activity Description
Khanh X Nguyen, Dipendra Misra, Robert Schapire, Miroslav Dudik, Patrick Shafto
We present a novel interactive learning protocol that enables training request-fulfilling agents by verbally describing their activities. Unlike imitation learning (IL), our protocol allows the teaching agent to provide feedback in a language that is most appropriate for them. Compared with reward in reinforcement learning (RL), the description feedback is richer and allows for improved sample complexity. We develop a probabilistic framework and an algorithm that practically implements our protocol. Empirical results in two challenging request-fulfilling problems demonstrate the strengths of our approach: compared with RL baselines, it is more sample-efficient; compared with IL baselines, it achieves competitive success rates without requiring the teaching agent to be able to demonstrate the desired behavior using the learning agent’s actions. Apart from empirical evaluation, we also provide theoretical guarantees for our algorithm under certain assumptions about the teacher and the environment.
****
Nonmyopic Multifidelity Acitve Search
Quan Nguyen, Arghavan Modiri, Roman Garnett
Active search is a learning paradigm where we seek to identify as many members of a rare, valuable class as possible given a labeling budget. Previous work on active search has assumed access to a faithful (and expensive) oracle reporting e

xperimental results. However, some settings offer access to cheaper surrogates such as computational simulation that may aid in the search. We propose a model of multifidelity active search, as well as a novel, computationally efficient policy for this setting that is motivated by state-of-the-art classical policies. Our policy is nonmyopic and budget aware, allowing for a dynamic tradeoff between exploration and exploitation. We evaluate the performance of our solution on real-world datasets and demonstrate significantly better performance than natural benchmarks.
****
Tight Bounds on the Smallest Eigenvalue of the Neural Tangent Kernel for Deep ReLU Networks
Quynh Nguyen, Marco Mondelli, Guido F Montufar
A recent line of work has analyzed the theoretical properties of deep neural networks via the Neural Tangent Kernel (NTK). In particular, the smallest eigenvalue of the NTK has been related to the memorization capacity, the global convergence of gradient descent algorithms and the generalization of deep nets. However, existing results either provide bounds in the two-layer setting or assume that the spectrum of the NTK matrices is bounded away from 0 for multi-layer networks. In this paper, we provide tight bounds on the smallest eigenvalue of NTK matrices for deep ReLU nets, both in the limiting case of infinite widths and for finite widths. In the finite-width setting, the network architectures we consider are fairly general: we require the existence of a wide layer with roughly order of N neurons, N being the number of data samples; and the scaling of the remaining layer widths is arbitrary (up to logarithmic factors). To obtain our results, we analyze various quantities of independent interest: we give lower bounds on the smallest singular value of hidden feature matrices, and upper bounds on the Lipschitz constant of input-output feature maps.
****
Temporal Predictive Coding For Model-Based Planning In Latent Space
Tung D Nguyen, Rui Shu, Tuan Pham, Hung Bui, Stefano Ermon
High-dimensional observations are a major challenge in the application of model-based reinforcement learning (MBRL) to real-world environments. To handle high-dimensional sensory inputs, existing approaches use representation learning to map high-dimensional observations into a lower-dimensional latent space that is more amenable to dynamics estimation and planning. In this work, we present an information-theoretic approach that employs temporal predictive coding to encode elements in the environment that can be predicted across time. Since this approach focuses on encoding temporally-predictable information, we implicitly prioritize the encoding of task-relevant components over nuisance information within the environment that are provably task-irrelevant. By learning this representation in conjunction with a recurrent state space model, we can then perform planning in latent space. We evaluate our model on a challenging modification of standard DMControl tasks where the background is replaced with natural videos that contain complex but irrelevant information to the planning task. Our experiments show that our model is superior to existing methods in the challenging complex-background setting while remaining competitive with current state-of-the-art models in the standard setting.
****
Differentially Private Densest Subgraph Detection
Dung Nguyen, Anil Vullikanti
Densest subgraph detection is a fundamental graph mining problem, with a large number of applications. There has been a lot of work on efficient algorithms for finding the densest subgraph in massive networks. However, in many domains, the network is private, and returning a densest subgraph can reveal information about the network. Differential privacy is a powerful framework to handle such settings. We study the densest subgraph problem in the edge privacy model, in which the edges of the graph are private. We present the first sequential and parallel differentially private algorithms for this problem. We show that our algorithms have an additive approximation guarantee. We evaluate our algorithms on a large number of real-world networks, and observe a good privacy-accuracy tradeoff when

the network has high density.
****
Data Augmentation for Meta-Learning
Renkun Ni, Micah Goldblum, Amr Sharaf, Kezhi Kong, Tom Goldstein
Conventional image classifiers are trained by randomly sampling mini-batches of images. To achieve state-of-the-art performance, practitioners use sophisticated data augmentation schemes to expand the amount of training data available for sampling. In contrast, meta-learning algorithms sample support data, query data, and tasks on each training step. In this complex sampling scenario, data augmentation can be used not only to expand the number of images available per class, but also to generate entirely new classes/tasks. We systematically dissect the meta-learning pipeline and investigate the distinct ways in which data augmentation can be integrated at both the image and class levels. Our proposed meta-specific data augmentation significantly improves the performance of meta-learners on few-shot classification benchmarks.
****
Improved Denoising Diffusion Probabilistic Models
Alexander Quinn Nichol, Prafulla Dhariwal
Denoising diffusion probabilistic models (DDPM) are a class of generative models which have recently been shown to produce excellent samples. We show that with a few simple modifications, DDPMs can also achieve competitive log-likelihoods while maintaining high sample quality. Additionally, we find that learning variances of the reverse diffusion process allows sampling with an order of magnitude fewer forward passes with a negligible difference in sample quality, which is important for the practical deployment of these models. We additionally use precision and recall to compare how well DDPMs and GANs cover the target distribution. Finally, we show that the sample quality and likelihood of these models scale smoothly with model capacity and training compute, making them easily scalable. We release our code and pre-trained models at https://github.com/openai/improved-diffusion.
****
Smooth p-Wasserstein Distance: Structure, Empirical Approximation, and Statistical Applications
Sloan Nietert, Ziv Goldfeld, Kengo Kato
Discrepancy measures between probability distributions, often termed statistical distances, are ubiquitous in probability theory, statistics and machine learning. To combat the curse of dimensionality when estimating these distances from data, recent work has proposed smoothing out local irregularities in the measured distributions via convolution with a Gaussian kernel. Motivated by the scalability of this framework to high dimensions, we investigate the structural and statistical behavior of the Gaussian-smoothed p-Wasserstein distance $\mathsf{W}_p^{(\sigma)}$, for arbitrary $p\geq 1$. After establishing basic metric and topological properties of $\mathsf{W}_p^{(\sigma)}$, we explore the asymptotic statistical properties of $\mathsf{W}_p^{(\sigma)}(\hat{\mu}_n,\mu)$, where $\hat{\mu}_n$ is the empirical distribution of n independent observations from μ. We prove that $\mathsf{W}_p^{(\sigma)}$ enjoys a parametric empirical convergence rate of $n^{-1/2}$, which contrasts the $n^{-1/d}$ rate for unsmoothed $\Wp$ when $d \geq 3$. Our proof relies on controlling $\mathsf{W}_p^{(\sigma)}$ by a pth-order smooth Sobolev distance $\mathsf{d}_p^{(\sigma)}$ and deriving the limit distribution of $\sqrt{n}\,\mathsf{d}_p^{(\sigma)}(\hat{\mu}_n,\mu)$ for all dimensions d. As applications, we provide asymptotic guarantees for two-sample testing and minimum distance estimation using $\mathsf{W}_p^{(\sigma)}$, with experiments for $p=2$ using a maximum mean discrepancy formulation of $\mathsf{d}_2^{(\sigma)}$.
****
AdaXpert: Adapting Neural Architecture for Growing Data
Shuaicheng Niu, Jiaxiang Wu, Guanghui Xu, Yifan Zhang, Yong Guo, Peilin Zhao, Peng Wang, Mingkui Tan
In real-world applications, data often come in a growing manner, where the data volume and the number of classes may increase dynamically. This will bring a cri

tical challenge for learning: given the increasing data volume or the number of classes, one has to instantaneously adjust the neural model capacity to obtain promising performance. Existing methods either ignore the growing nature of data or seek to independently search an optimal architecture for a given dataset, and thus are incapable of promptly adjusting the architectures for the changed data. To address this, we present a neural architecture adaptation method, namely Adaptation eXpert (AdaXpert), to efficiently adjust previous architectures on the growing data. Specifically, we introduce an architecture adjuster to generate a suitable architecture for each data snapshot, based on the previous architecture and the different extent between current and previous data distributions. Furthermore, we propose an adaptation condition to determine the necessity of adjustment, thereby avoiding unnecessary and time-consuming adjustments. Extensive experiments on two growth scenarios (increasing data volume and number of classes) demonstrate the effectiveness of the proposed method.
****
Asynchronous Decentralized Optimization With Implicit Stochastic Variance Reduction
Kenta Niwa, Guoqiang Zhang, W. Bastiaan Kleijn, Noboru Harada, Hiroshi Sawada, Akinori Fujino
A novel asynchronous decentralized optimization method that follows Stochastic Variance Reduction (SVR) is proposed. Average consensus algorithms, such as Decentralized Stochastic Gradient Descent (DSGD), facilitate distributed training of machine learning models. However, the gradient will drift within the local nodes due to statistical heterogeneity of the subsets of data residing on the nodes and long communication intervals. To overcome the drift problem, (i) Gradient Tracking-SVR (GT-SVR) integrates SVR into DSGD and (ii) Edge-Consensus Learning (ECL) solves a model constrained minimization problem using a primal-dual formalism. In this paper, we reformulate the update procedure of ECL such that it implicitly includes the gradient modification of SVR by optimally selecting a constraint-strength control parameter. Through convergence analysis and experiments, we confirmed that the proposed ECL with Implicit SVR (ECL-ISVR) is stable and approximately reaches the reference performance obtained with computation on a single-node using full data set.
****
WGAN with an Infinitely Wide Generator Has No Spurious Stationary Points
Albert No, Taeho Yoon, Kwon Sehyun, Ernest K Ryu
Generative adversarial networks (GAN) are a widely used class of deep generative models, but their minimax training dynamics are not understood very well. In this work, we show that GANs with a 2-layer infinite-width generator and a 2-layer finite-width discriminator trained with stochastic gradient ascent-descent have no spurious stationary points. We then show that when the width of the generator is finite but wide, there are no spurious stationary points within a ball whose radius becomes arbitrarily large (to cover the entire parameter space) as the width goes to infinity.
****
The Impact of Record Linkage on Learning from Feature Partitioned Data
Richard Nock, Stephen Hardy, Wilko Henecka, Hamish Ivey-Law, Jakub Nabaglo, Giorgio Patrini, Guillaume Smith, Brian Thorne
There has been recently a significant boost to machine learning with distributed data, in particular with the success of federated learning. A common and very challenging setting is that of vertical or feature partitioned data, when multiple data providers hold different features about common entities. In general, training needs to be preceded by record linkage (RL), a step that finds the correspondence between the observations of the datasets. RL is prone to mistakes in the real world. Despite the importance of the problem, there has been so far no formal assessment of the way in which RL errors impact learning models. Work in the area either use heuristics or assume that the optimal RL is known in advance. In this paper, we provide the first assessment of the problem for supervised learning. For wide sets of losses, we provide technical conditions under which the classifier learned after noisy RL converges (with the data size) to the best class

ifier that would be learned from mistake-free RL. This yields new insights on th
e way the pipeline RL + ML operates, from the role of large margin classificatio
n on dampening the impact of RL mistakes to clues on how to further optimize RL
as a preprocessing step to ML. Experiments on a large UCI benchmark validate tho
se formal observations.
****
Accuracy, Interpretability, and Differential Privacy via Explainable Boosting
Harsha Nori, Rich Caruana, Zhiqi Bu, Judy Hanwen Shen, Janardhan Kulkarni
We show that adding differential privacy to Explainable Boosting Machines (EBMs)
, a recent method for training interpretable ML models, yields state-of-the-art
accuracy while protecting privacy. Our experiments on multiple classification an
d regression datasets show that DP-EBM models suffer surprisingly little accurac
y loss even with strong differential privacy guarantees. In addition to high acc
uracy, two other benefits of applying DP to EBMs are: a) trained models provide
exact global and local interpretability, which is often important in settings wh
ere differential privacy is needed; and b) the models can be edited after traini
ng without loss of privacy to correct errors which DP noise may have introduced.
****
Posterior Value Functions: Hindsight Baselines for Policy Gradient Methods
Chris Nota, Philip Thomas, Bruno C. Da Silva
Hindsight allows reinforcement learning agents to leverage new observations to m
ake inferences about earlier states and transitions. In this paper, we exploit t
he idea of hindsight and introduce posterior value functions. Posterior value fu
nctions are computed by inferring the posterior distribution over hidden compone
nts of the state in previous timesteps and can be used to construct novel unbias
ed baselines for policy gradient methods. Importantly, we prove that these basel
ines reduce (and never increase) the variance of policy gradient estimators comp
ared to traditional state value functions. While the posterior value function is
 motivated by partial observability, we extend these results to arbitrary stocha
stic MDPs by showing that hindsight-capable agents can model stochasticity in th
e environment as a special case of partial observability. Finally, we introduce
a pair of methods for learning posterior value functions and prove their converg
ence.
****
Global inducing point variational posteriors for Bayesian neural networks and de
ep Gaussian processes
Sebastian W Ober, Laurence Aitchison
We consider the optimal approximate posterior over the top-layer weights in a Ba
yesian neural network for regression, and show that it exhibits strong dependenc
ies on the lower-layer weights. We adapt this result to develop a correlated app
roximate posterior over the weights at all layers in a Bayesian neural network.
We extend this approach to deep Gaussian processes, unifying inference in the tw
o model classes. Our approximate posterior uses learned "global” inducing points
, which are defined only at the input layer and propagated through the network t
o obtain inducing inputs at subsequent layers. By contrast, standard, "local”, i
nducing point methods from the deep Gaussian process literature optimise a separ
ate set of inducing inputs at every layer, and thus do not model correlations ac
ross layers. Our method gives state-of-the-art performance for a variational Bay
esian method, without data augmentation or tempering, on CIFAR-10 of 86.7%, whic
h is comparable to SGMCMC without tempering but with data augmentation (88% in W
enzel et al. 2020).
****
Regularizing towards Causal Invariance: Linear Models with Proxies
Michael Oberst, Nikolaj Thams, Jonas Peters, David Sontag
We propose a method for learning linear models whose predictive performance is r
obust to causal interventions on unobserved variables, when noisy proxies of tho
se variables are available. Our approach takes the form of a regularization term
 that trades off between in-distribution performance and robustness to intervent
ions. Under the assumption of a linear structural causal model, we show that a s
ingle proxy can be used to create estimators that are prediction optimal under i

nterventions of bounded strength. This strength depends on the magnitude of the measurement noise in the proxy, which is, in general, not identifiable. In the case of two proxy variables, we propose a modified estimator that is prediction optimal under interventions up to a known strength. We further show how to extend these estimators to scenarios where additional information about the "test time" intervention is available during training. We evaluate our theoretical findings in synthetic experiments and using real data of hourly pollution levels across several cities in China.
****
Sparsity-Agnostic Lasso Bandit
Min-Hwan Oh, Garud Iyengar, Assaf Zeevi
We consider a stochastic contextual bandit problem where the dimension d of the feature vectors is potentially large, however, only a sparse subset of features of cardinality $s_0 \ll d$ affect the reward function. Essentially all existing algorithms for sparse bandits require a priori knowledge of the value of the sparsity index s_0. This knowledge is almost never available in practice, and misspecification of this parameter can lead to severe deterioration in the performance of existing methods. The main contribution of this paper is to propose an algorithm that does not require prior knowledge of the sparsity index s_0 and establish tight regret bounds on its performance under mild conditions. We also comprehensively evaluate our proposed algorithm numerically and show that it consistently outperforms existing methods, even when the correct sparsity index is revealed to them but is kept hidden from our algorithm.
****
Autoencoder Image Interpolation by Shaping the Latent Space
Alon Oring, Zohar Yakhini, Yacov Hel-Or
One of the fascinating properties of deep learning is the ability of the network to reveal the underlying factors characterizing elements in datasets of different types. Autoencoders represent an effective approach for computing these factors. Autoencoders have been studied in the context of enabling interpolation between data points by decoding convex combinations of latent vectors. However, this interpolation often leads to artifacts or produces unrealistic results during reconstruction. We argue that these incongruities are due to the structure of the latent space and to the fact that such naively interpolated latent vectors deviate from the data manifold. In this paper, we propose a regularization technique that shapes the latent representation to follow a manifold that is consistent with the training images and that forces the manifold to be smooth and locally convex. This regularization not only enables faithful interpolation between data points, as we show herein but can also be used as a general regularization technique to avoid overfitting or to produce new samples for data augmentation.
****
Generalization Guarantees for Neural Architecture Search with Train-Validation Split
Samet Oymak, Mingchen Li, Mahdi Soltanolkotabi
Neural Architecture Search (NAS) is a popular method for automatically designing optimized deep-learning architectures. NAS methods commonly use bilevel optimization where one optimizes the weights over the training data (lower-level problem) and hyperparameters - such as the architecture - over the validation data (upper-level problem). This paper explores the statistical aspects of such problems with train-validation splits. In practice, the lower-level problem is often overparameterized and can easily achieve zero loss. Thus, a-priori, it seems impossible to distinguish the right hyperparameters based on training loss alone which motivates a better understanding of train-validation split. To this aim, we first show that refined properties of the validation loss such as risk and hyper-gradients are indicative of those of the true test loss and help prevent overfitting with a near-minimal validation sample size. Importantly, this is established for continuous search spaces which are relevant for differentiable search schemes. We then establish generalization bounds for NAS problems with an emphasis on an activation search problem and gradient-based methods. Finally, we show rigorous connections between NAS and low-rank matrix learning which leads to algorithm

ic insights where the solution of the upper problem can be accurately learned vi
a spectral methods to achieve near-minimal risk.
****
Vector Quantized Models for Planning
Sherjil Ozair, Yazhe Li, Ali Razavi, Ioannis Antonoglou, Aaron Van Den Oord, Ori
ol Vinyals
Recent developments in the field of model-based RL have proven successful in a r
ange of environments, especially ones where planning is essential. However, such
 successes have been limited to deterministic fully-observed environments. We pr
esent a new approach that handles stochastic and partially-observable environmen
ts. Our key insight is to use discrete autoencoders to capture the multiple poss
ible effects of an action in a stochastic environment. We use a stochastic varia
nt of Monte Carlo tree search to plan over both the agent's actions and the disc
rete latent variables representing the environment's response. Our approach sign
ificantly outperforms an offline version of MuZero on a stochastic interpretatio
n of chess where the opponent is considered part of the environment. We also sho
w that our approach scales to DeepMind Lab, a first-person 3D environment with l
arge visual observations and partial observability.
****
Training Adversarially Robust Sparse Networks via Bayesian Connectivity Sampling
Ozan Özdenizci, Robert Legenstein
Deep neural networks have been shown to be susceptible to adversarial attacks. T
his lack of adversarial robustness is even more pronounced when models are compr
essed in order to meet hardware limitations. Hence, if adversarial robustness is
 an issue, training of sparsely connected networks necessitates considering adve
rsarially robust sparse learning. Motivated by the efficient and stable computat
ional function of the brain in the presence of a highly dynamic synaptic connect
ivity structure, we propose an intrinsically sparse rewiring approach to train n
eural networks with state-of-the-art robust learning objectives under high spars
ity. Importantly, in contrast to previously proposed pruning techniques, our app
roach satisfies global connectivity constraints throughout robust optimization,
i.e., it does not require dense pre-training followed by pruning. Based on a Bay
esian posterior sampling principle, a network rewiring process simultaneously le
arns the sparse connectivity structure and the robustness-accuracy trade-off bas
ed on the adversarial learning objective. Although our networks are sparsely con
nected throughout the whole training process, our experimental benchmark evaluat
ions show that their performance is superior to recently proposed robustness-awa
re network pruning methods which start from densely connected networks.
****
Opening the Blackbox: Accelerating Neural Differential Equations by Regularizing
 Internal Solver Heuristics
Avik Pal, Yingbo Ma, Viral Shah, Christopher V Rackauckas
Democratization of machine learning requires architectures that automatically ad
apt to new problems. Neural Differential Equations (NDEs) have emerged as a popu
lar modeling framework by removing the need for ML practitioners to choose the n
umber of layers in a recurrent model. While we can control the computational cos
t by choosing the number of layers in standard architectures, in NDEs the number
 of neural network evaluations for a forward pass can depend on the number of st
eps of the adaptive ODE solver. But, can we force the NDE to learn the version w
ith the least steps while not increasing the training cost? Current strategies t
o overcome slow prediction require high order automatic differentiation, leading
 to significantly higher training time. We describe a novel regularization metho
d that uses the internal cost heuristics of adaptive differential equation solve
rs combined with discrete adjoint sensitivities to guide the training process to
wards learning NDEs that are easier to solve. This approach opens up the blackbo
x numerical analysis behind the differential equation solver's algorithm and dir
ectly uses its local error estimates and stiffness heuristics as cheap and accur
ate cost estimates. We incorporate our method without any change in the underlyi
ng NDE framework and show that our method extends beyond Ordinary Differential E
quations to accommodate Neural Stochastic Differential Equations. We demonstrate

 how our approach can halve the prediction time and, unlike other methods which can increase the training time by an order of magnitude, we demonstrate similar reduction in training times. Together this showcases how the knowledge embedded within state-of-the-art equation solvers can be used to enhance machine learning
.
****
RNN with Particle Flow for Probabilistic Spatio-temporal Forecasting
Soumyasundar Pal, Liheng Ma, Yingxue Zhang, Mark Coates
Spatio-temporal forecasting has numerous applications in analyzing wireless, tra
ffic, and financial networks. Many classical statistical models often fall short
 in handling the complexity and high non-linearity present in time-series data.
Recent advances in deep learning allow for better modelling of spatial and tempo
ral dependencies. While most of these models focus on obtaining accurate point f
orecasts, they do not characterize the prediction uncertainty. In this work, we
consider the time-series data as a random realization from a nonlinear state-spa
ce model and target Bayesian inference of the hidden states for probabilistic fo
recasting. We use particle flow as the tool for approximating the posterior dist
ribution of the states, as it is shown to be highly effective in complex, high-d
imensional settings. Thorough experimentation on several real world time-series
datasets demonstrates that our approach provides better characterization of unce
rtainty while maintaining comparable accuracy to the state-of-the-art point fore
casting methods.
****
Inference for Network Regression Models with Community Structure
Mengjie Pan, Tyler Mccormick, Bailey Fosdick
Network regression models, where the outcome comprises the valued edge in a netw
ork and the predictors are actor or dyad-level covariates, are used extensively
in the social and biological sciences. Valid inference relies on accurately mode
ling the residual dependencies among the relations. Frequently homogeneity assum
ptions are placed on the errors which are commonly incorrect and ignore critical
 natural clustering of the actors. In this work, we present a novel regression m
odeling framework that models the errors as resulting from a community-based dep
endence structure and exploits the subsequent exchangeability properties of the
error distribution to obtain parsimonious standard errors for regression paramet
ers.
****
Latent Space Energy-Based Model of Symbol-Vector Coupling for Text Generation an
d Classification
Bo Pang, Ying Nian Wu
We propose a latent space energy-based prior model for text generation and class
ification. The model stands on a generator network that generates the text seque
nce based on a continuous latent vector. The energy term of the prior model coup
les a continuous latent vector and a symbolic one-hot vector, so that discrete c
ategory can be inferred from the observed example based on the continuous latent
 vector. Such a latent space coupling naturally enables incorporation of informa
tion bottleneck regularization to encourage the continuous latent vector to extr
act information from the observed example that is informative of the underlying
category. In our learning method, the symbol-vector coupling, the generator netw
ork and the inference network are learned jointly. Our model can be learned in a
n unsupervised setting where no category labels are provided. It can also be lea
rned in semi-supervised setting where category labels are provided for a subset
of training examples. Our experiments demonstrate that the proposed model learns
 well-structured and meaningful latent space, which (1) guides the generator to
generate text with high quality, diversity, and interpretability, and (2) effect
ively classifies text.
****
Leveraging Good Representations in Linear Contextual Bandits
Matteo Papini, Andrea Tirinzoni, Marcello Restelli, Alessandro Lazaric, Matteo P
irotta
The linear contextual bandit literature is mostly focused on the design of effic

ient learning algorithms for a given representation. However, a contextual bandit problem may admit multiple linear representations, each one with different characteristics that directly impact the regret of the learning algorithm. In particular, recent works showed that there exist “good” representations for which constant problem-dependent regret can be achieved. In this paper, we first provide a systematic analysis of the different definitions of “good” representations proposed in the literature. We then propose a novel selection algorithm able to adapt to the best representation in a set of M candidates. We show that the regret is indeed never worse than the regret obtained by running \textsc{LinUCB} on best representation (up to a $\ln M$ factor). As a result, our algorithm achieves constant regret if a “good” representation is available in the set. Furthermore, we show the algorithm may still achieve constant regret by implicitly constructing a “good” representation, even when none of the initial representations is “good”. Finally, we validate our theoretical findings in a number of standard contextual bandit problems.
****
Wasserstein Distributional Normalization For Robust Distributional Certification of Noisy Labeled Data
Sung Woo Park, Junseok Kwon
We propose a novel Wasserstein distributional normalization method that can classify noisy labeled data accurately. Recently, noisy labels have been successfully handled based on small-loss criteria, but have not been clearly understood from the theoretical point of view. In this paper, we address this problem by adopting distributionally robust optimization (DRO). In particular, we present a theoretical investigation of the distributional relationship between uncertain and certain samples based on the small-loss criteria. Our method takes advantage of this relationship to exploit useful information from uncertain samples. To this end, we normalize uncertain samples into the robustly certified region by introducing the non-parametric Ornstein-Ulenbeck type of Wasserstein gradient flows called Wasserstein distributional normalization, which is cheap and fast to implement. We verify that network confidence and distributional certification are fundamentally correlated and show the concentration inequality when the network escapes from over-parameterization. Experimental results demonstrate that our non-parametric classification method outperforms other parametric baselines on the Clothing1M and CIFAR-10/100 datasets when the data have diverse noisy labels.
****
Unsupervised Representation Learning via Neural Activation Coding
Yookoon Park, Sangho Lee, Gunhee Kim, David Blei
We present neural activation coding (NAC) as a novel approach for learning deep representations from unlabeled data for downstream applications. We argue that the deep encoder should maximize its nonlinear expressivity on the data for downstream predictors to take full advantage of its representation power. To this end, NAC maximizes the mutual information between activation patterns of the encoder and the data over a noisy communication channel. We show that learning for a noise-robust activation code increases the number of distinct linear regions of ReLU encoders, hence the maximum nonlinear expressivity. More interestingly, NAC learns both continuous and discrete representations of data, which we respectively evaluate on two downstream tasks: (i) linear classification on CIFAR-10 and ImageNet-1K and (ii) nearest neighbor retrieval on CIFAR-10 and FLICKR-25K. Empirical results show that NAC attains better or comparable performance on both tasks over recent baselines including SimCLR and DistillHash. In addition, NAC pretraining provides significant benefits to the training of deep generative models. Our code is available at https://github.com/yookoon/nac.
****
Conditional Distributional Treatment Effect with Kernel Conditional Mean Embeddings and U-Statistic Regression
Junhyung Park, Uri Shalit, Bernhard Schölkopf, Krikamol Muandet
We propose to analyse the conditional distributional treatment effect (CoDiTE), which, in contrast to the more common conditional average treatment effect (CATE), is designed to encode a treatment’s distributional aspects beyond the mean. W

e first introduce a formal definition of the CoDiTE associated with a distance f
unction between probability measures. Then we discuss the CoDiTE associated with
 the maximum mean discrepancy via kernel conditional mean embeddings, which, cou
pled with a hypothesis test, tells us whether there is any conditional distribut
ional effect of the treatment. Finally, we investigate what kind of conditional
distributional effect the treatment has, both in an exploratory manner via the c
onditional witness function, and in a quantitative manner via U-statistic regres
sion, generalising the CATE to higher-order moments. Experiments on synthetic, s
emi-synthetic and real datasets demonstrate the merits of our approach.
****
Generative Adversarial Networks for Markovian Temporal Dynamics: Stochastic Cont
inuous Data Generation
Sung Woo Park, Dong Wook Shu, Junseok Kwon
In this paper, we present a novel generative adversarial network (GAN) that can
describe Markovian temporal dynamics. To generate stochastic sequential data, we
 introduce a novel stochastic differential equation-based conditional generator
and spatial-temporal constrained discriminator networks. To stabilize the learni
ng dynamics of the min-max type of the GAN objective function, we propose well-p
osed constraint terms for both networks. We also propose a novel conditional Mar
kov Wasserstein distance to induce a pathwise Wasserstein distance. The experime
ntal results demonstrate that our method outperforms state-of-the-art methods us
ing several different types of data.
****
Optimal Counterfactual Explanations in Tree Ensembles
Axel Parmentier, Thibaut Vidal
Counterfactual explanations are usually generated through heuristics that are se
nsitive to the search’s initial conditions. The absence of guarantees of perform
ance and robustness hinders trustworthiness. In this paper, we take a discipline
d approach towards counterfactual explanations for tree ensembles. We advocate f
or a model-based search aiming at "optimal" explanations and propose efficient m
ixed-integer programming approaches. We show that isolation forests can be model
ed within our framework to focus the search on plausible explanations with a low
 outlier score. We provide comprehensive coverage of additional constraints that
 model important objectives, heterogeneous data types, structural constraints on
 the feature space, along with resource and actionability restrictions. Our expe
rimental analyses demonstrate that the proposed search approach requires a compu
tational effort that is orders of magnitude smaller than previous mathematical p
rogramming algorithms. It scales up to large data sets and tree ensembles, where
 it provides, within seconds, systematic explanations grounded on well-defined m
odels solved to optimality.
****
PHEW : Constructing Sparse Networks that Learn Fast and Generalize Well without
Training Data
Shreyas Malakarjun Patil, Constantine Dovrolis
Methods that sparsify a network at initialization are important in practice beca
use they greatly improve the efficiency of both learning and inference. Our work
 is based on a recently proposed decomposition of the Neural Tangent Kernel (NTK
) that has decoupled the dynamics of the training process into a data-dependent
component and an architecture-dependent kernel {–} the latter referred to as Pat
h Kernel. That work has shown how to design sparse neural networks for faster co
nvergence, without any training data, using the Synflow-L2 algorithm. We first s
how that even though Synflow-L2 is optimal in terms of convergence, for a given
network density, it results in sub-networks with “bottleneck” (narrow) layers {–
} leading to poor performance as compared to other data-agnostic methods that us
e the same number of parameters. Then we propose a new method to construct spars
e networks, without any training data, referred to as Paths with Higher-Edge Wei
ghts (PHEW). PHEW is a probabilistic network formation method based on biased ra
ndom walks that only depends on the initial weights. It has similar path kernel
properties as Synflow-L2 but it generates much wider layers, resulting in better
 generalization and performance. PHEW achieves significant improvements over the

data-independent SynFlow and SynFlow-L2 methods at a wide range of network dens
ities.
****
CombOptNet: Fit the Right NP-Hard Problem by Learning Integer Programming Constr
aints
Anselm Paulus, Michal Rolinek, Vit Musil, Brandon Amos, Georg Martius
Bridging logical and algorithmic reasoning with modern machine learning techniqu
es is a fundamental challenge with potentially transformative impact. On the alg
orithmic side, many NP-hard problems can be expressed as integer programs, in wh
ich the constraints play the role of their ’combinatorial specification’. In thi
s work, we aim to integrate integer programming solvers into neural network arch
itectures as layers capable of learning both the cost terms and the constraints.
 The resulting end-to-end trainable architectures jointly extract features from
raw data and solve a suitable (learned) combinatorial problem with state-of-the-
art integer programming solvers. We demonstrate the potential of such layers wit
h an extensive performance analysis on synthetic data and with a demonstration o
n a competitive computer vision keypoint matching benchmark.
****
Ensemble Bootstrapping for Q-Learning
Oren Peer, Chen Tessler, Nadav Merlis, Ron Meir
Q-learning (QL), a common reinforcement learning algorithm, suffers from over-es
timation bias due to the maximization term in the optimal Bellman operator. This
 bias may lead to sub-optimal behavior. Double-Q-learning tackles this issue by
utilizing two estimators, yet results in an under-estimation bias. Similar to ov
er-estimation in Q-learning, in certain scenarios, the under-estimation bias may
 degrade performance. In this work, we introduce a new bias-reduced algorithm ca
lled Ensemble Bootstrapped Q-Learning (EBQL), a natural extension of Double-Q-le
arning to ensembles. We analyze our method both theoretically and empirically. T
heoretically, we prove that EBQL-like updates yield lower MSE when estimating th
e maximal mean of a set of independent random variables. Empirically, we show th
at there exist domains where both over and under-estimation result in sub-optima
l performance. Finally, We demonstrate the superior performance of a deep RL var
iant of EBQL over other deep QL algorithms for a suite of ATARI games.
****
Homomorphic Sensing: Sparsity and Noise
Liangzu Peng, Boshi Wang, Manolis Tsakiris
\emph{Unlabeled sensing} is a recent problem encompassing many data science and
engineering applications and typically formulated as solving linear equations wh
ose right-hand side vector has undergone an unknown permutation. It was generali
zed to the \emph{homomorphic sensing} problem by replacing the unknown permutati
on with an unknown linear map from a given finite set of linear maps. In this pa
per we present tighter and simpler conditions for the homomorphic sensing proble
m to admit a unique solution. We show that this solution is locally stable under
 noise, while under a sparsity assumption it remains unique under less demanding
 conditions. Sparsity in the context of unlabeled sensing leads to the problem o
f \textit{unlabeled compressed sensing}, and a consequence of our general theory
 is the existence under mild conditions of a unique sparsest solution. On the al
gorithmic level, we solve unlabeled compressed sensing by an iterative algorithm
 validated by synthetic data experiments. Finally, under the unifying homomorphi
c sensing framework we connect unlabeled sensing to other important practical pr
oblems.
****
How could Neural Networks understand Programs?
Dinglan Peng, Shuxin Zheng, Yatao Li, Guolin Ke, Di He, Tie-Yan Liu
Semantic understanding of programs is a fundamental problem for programming lang
uage processing (PLP). Recent works that learn representations of code based on
pre-training techniques in NLP have pushed the frontiers in this direction. Howe
ver, the semantics of PL and NL have essential differences. These being ignored,
 we believe it is difficult to build a model to better understand programs, by e
ither directly applying off-the-shelf NLP pre-training techniques to the source

code, or adding features to the model by the heuristic. In fact, the semantics of a program can be rigorously defined by formal semantics in PL theory. For example, the operational semantics, describes the meaning of a valid program as updating the environment (i.e., the memory address-value function) through fundamental operations, such as memory I/O and conditional branching. Inspired by this, we propose a novel program semantics learning paradigm, that the model should learn from information composed of (1) the representations which align well with the fundamental operations in operational semantics, and (2) the information of environment transition, which is indispensable for program understanding. To validate our proposal, we present a hierarchical Transformer-based pre-training model called OSCAR to better facilitate the understanding of programs. OSCAR learns from intermediate representation (IR) and an encoded representation derived from static analysis, which are used for representing the fundamental operations and approximating the environment transitions respectively. OSCAR empirically shows the outstanding capability of program semantics understanding on many practical software engineering tasks. Code and models are released at: \url{https://github.com/pdlan/OSCAR}.
****
Privacy-Preserving Video Classification with Convolutional Neural Networks
Sikha Pentyala, Rafael Dowsley, Martine De Cock
Many video classification applications require access to personal data, thereby posing an invasive security risk to the users' privacy. We propose a privacy-preserving implementation of single-frame method based video classification with convolutional neural networks that allows a party to infer a label from a video without necessitating the video owner to disclose their video to other entities in an unencrypted manner. Similarly, our approach removes the requirement of the classifier owner from revealing their model parameters to outside entities in plaintext. To this end, we combine existing Secure Multi-Party Computation (MPC) protocols for private image classification with our novel MPC protocols for oblivious single-frame selection and secure label aggregation across frames. The result is an end-to-end privacy-preserving video classification pipeline. We evaluate our proposed solution in an application for private human emotion recognition. Our results across a variety of security settings, spanning honest and dishonest majority configurations of the computing parties, and for both passive and active adversaries, demonstrate that videos can be classified with state-of-the-art accuracy, and without leaking sensitive user information.
****
Rissanen Data Analysis: Examining Dataset Characteristics via Description Length
Ethan Perez, Douwe Kiela, Kyunghyun Cho
We introduce a method to determine if a certain capability helps to achieve an accurate model of given data. We view labels as being generated from the inputs by a program composed of subroutines with different capabilities, and we posit that a subroutine is useful if and only if the minimal program that invokes it is shorter than the one that does not. Since minimum program length is uncomputable, we instead estimate the labels' minimum description length (MDL) as a proxy, giving us a theoretically-grounded method for analyzing dataset characteristics. We call the method Rissanen Data Analysis (RDA) after the father of MDL, and we showcase its applicability on a wide variety of settings in NLP, ranging from evaluating the utility of generating subquestions before answering a question, to analyzing the value of rationales and explanations, to investigating the importance of different parts of speech, and uncovering dataset gender bias.
****
Modelling Behavioural Diversity for Learning in Open-Ended Games
Nicolas Perez-Nieves, Yaodong Yang, Oliver Slumbers, David H Mguni, Ying Wen, Jun Wang
Promoting behavioural diversity is critical for solving games with non-transitive dynamics where strategic cycles exist, and there is no consistent winner (e.g., Rock-Paper-Scissors). Yet, there is a lack of rigorous treatment for defining diversity and constructing diversity-aware learning dynamics. In this work, we offer a geometric interpretation of behavioural diversity in games and introduce

a novel diversity metric based on \emph{determinantal point processes} (DPP). By incorporating the diversity metric into best-response dynamics, we develop \emph{diverse fictitious play} and \emph{diverse policy-space response oracle} for solving normal-form games and open-ended games. We prove the uniqueness of the diverse best response and the convergence of our algorithms on two-player games. Importantly, we show that maximising the DPP-based diversity metric guarantees to enlarge the \emph{gamescape} – convex polytopes spanned by agents' mixtures of strategies. To validate our diversity-aware solvers, we test on tens of games that show strong non-transitivity. Results suggest that our methods achieve at least the same, and in most games, lower exploitability than PSRO solvers by finding effective and diverse strategies.
****
From Poincaré Recurrence to Convergence in Imperfect Information Games: Finding Equilibrium via Regularization
Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, Georgios Piliouras, Marc Lanctot, Karl Tuyls
In this paper we investigate the Follow the Regularized Leader dynamics in sequential imperfect information games (IIG). We generalize existing results of Poincar{é} recurrence from normal-form games to zero-sum two-player imperfect information games and other sequential game settings. We then investigate how adapting the reward (by adding a regularization term) of the game can give strong convergence guarantees in monotone games. We continue by showing how this reward adaptation technique can be leveraged to build algorithms that converge exactly to the Nash equilibrium. Finally, we show how these insights can be directly used to build state-of-the-art model-free algorithms for zero-sum two-player Imperfect Information Games (IIG).
****
Spectral Smoothing Unveils Phase Transitions in Hierarchical Variational Autoencoders
Adeel Pervez, Efstratios Gavves
Variational autoencoders with deep hierarchies of stochastic layers have been known to suffer from the problem of posterior collapse, where the top layers fall back to the prior and become independent of input. We suggest that the hierarchical VAE objective explicitly includes the variance of the function parameterizing the mean and variance of the latent Gaussian distribution which itself is often a high variance function. Building on this we generalize VAE neural networks by incorporating a smoothing parameter motivated by Gaussian analysis to reduce higher frequency components and consequently the variance in parameterizing functions and show that this can help to solve the problem of posterior collapse. We further show that under such smoothing the VAE loss exhibits a phase transition, where the top layer KL divergence sharply drops to zero at a critical value of the smoothing parameter that is similar for the same model across datasets. We validate the phenomenon across model configurations and datasets.
****
Differentiable Sorting Networks for Scalable Sorting and Ranking Supervision
Felix Petersen, Christian Borgelt, Hilde Kuehne, Oliver Deussen
Sorting and ranking supervision is a method for training neural networks end-to-end based on ordering constraints. That is, the ground truth order of sets of samples is known, while their absolute values remain unsupervised. For that, we propose differentiable sorting networks by relaxing their pairwise conditional swap operations. To address the problems of vanishing gradients and extensive blurring that arise with larger numbers of layers, we propose mapping activations to regions with moderate gradients. We consider odd-even as well as bitonic sorting networks, which outperform existing relaxations of the sorting operation. We show that bitonic sorting networks can achieve stable training on large input sets of up to 1024 elements.
****
Megaverse: Simulating Embodied Agents at One Million Experiences per Second
Aleksei Petrenko, Erik Wijmans, Brennan Shacklett, Vladlen Koltun

We present Megaverse, a new 3D simulation platform for reinforcement learning an
d embodied AI research. The efficient design of our engine enables physics-based
 simulation with high-dimensional egocentric observations at more than 1,000,000
 actions per second on a single 8-GPU node. Megaverse is up to 70x faster than D
eepMind Lab in fully-shaded 3D scenes with interactive objects. We achieve this
high simulation performance by leveraging batched simulation, thereby taking ful
l advantage of the massive parallelism of modern GPUs. We use Megaverse to build
 a new benchmark that consists of several single-agent and multi-agent tasks cov
ering a variety of cognitive challenges. We evaluate model-free RL on this bench
mark to provide baselines and facilitate future research.
****
Towards Practical Mean Bounds for Small Samples
My Phan, Philip Thomas, Erik Learned-Miller
Historically, to bound the mean for small sample sizes, practitioners have had t
o choose between using methods with unrealistic assumptions about the unknown di
stribution (e.g., Gaussianity) and methods like Hoeffding’s inequality that use
weaker assumptions but produce much looser (wider) intervals. In 1969, \citet{An
derson1969} proposed a mean confidence interval strictly better than or equal to
 Hoeffding’s whose only assumption is that the distribution’s support is contain
ed in an interval $[a,b]$. For the first time since then, we present a new famil
y of bounds that compares favorably to Anderson’s. We prove that each bound in t
he family has {\em guaranteed coverage}, i.e., it holds with probability at leas
t $1-\alpha$ for all distributions on an interval $[a,b]$. Furthermore, one of t
he bounds is tighter than or equal to Anderson’s for all samples. In simulations
, we show that for many distributions, the gain over Anderson’s bound is substan
tial.
****
DG-LMC: A Turn-key and Scalable Synchronous Distributed MCMC Algorithm via Lange
vin Monte Carlo within Gibbs
Vincent Plassier, Maxime Vono, Alain Durmus, Eric Moulines
Performing reliable Bayesian inference on a big data scale is becoming a keyston
e in the modern era of machine learning. A workhorse class of methods to achieve
 this task are Markov chain Monte Carlo (MCMC) algorithms and their design to ha
ndle distributed datasets has been the subject of many works. However, existing
methods are not completely either reliable or computationally efficient. In this
 paper, we propose to fill this gap in the case where the dataset is partitioned
 and stored on computing nodes within a cluster under a master/slaves architectu
re. We derive a user-friendly centralised distributed MCMC algorithm with provab
le scaling in high-dimensional settings. We illustrate the relevance of the prop
osed methodology on both synthetic and real data experiments.
****
GeomCA: Geometric Evaluation of Data Representations
Petra Poklukar, Anastasiia Varava, Danica Kragic
Evaluating the quality of learned representations without relying on a downstrea
m task remains one of the challenges in representation learning. In this work, w
e present Geometric Component Analysis (GeomCA) algorithm that evaluates represe
ntation spaces based on their geometric and topological properties. GeomCA can b
e applied to representations of any dimension, independently of the model that g
enerated them. We demonstrate its applicability by analyzing representations obt
ained from a variety of scenarios, such as contrastive learning models, generati
ve models and supervised learning models.
****
Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech
Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Kudinov
Recently, denoising diffusion probabilistic models and generative score matching
 have shown high potential in modelling complex data distributions while stochas
tic calculus has provided a unified point of view on these techniques allowing f
or flexible inference schemes. In this paper we introduce Grad-TTS, a novel text
-to-speech model with score-based decoder producing mel-spectrograms by graduall
y transforming noise predicted by encoder and aligned with text input by means o

f Monotonic Alignment Search. The framework of stochastic differential equations helps us to generalize conventional diffusion probabilistic models to the case of reconstructing data from noise with different parameters and allows to make this reconstruction flexible by explicitly controlling trade-off between sound quality and inference speed. Subjective human evaluation shows that Grad-TTS is competitive with state-of-the-art text-to-speech approaches in terms of Mean Opinion Score.

****

Bias-Free Scalable Gaussian Processes via Randomized Truncations

Andres Potapczynski, Luhuan Wu, Dan Biderman, Geoff Pleiss, John P Cunningham

Scalable Gaussian Process methods are computationally attractive, yet introduce modeling biases that require rigorous study. This paper analyzes two common techniques: early truncated conjugate gradients (CG) and random Fourier features (RFF). We find that both methods introduce a systematic bias on the learned hyperparameters: CG tends to underfit while RFF tends to overfit. We address these issues using randomized truncation estimators that eliminate bias in exchange for increased variance. In the case of RFF, we show that the bias-to-variance conversion is indeed a trade-off: the additional variance proves detrimental to optimization. However, in the case of CG, our unbiased learning procedure meaningfully outperforms its biased counterpart with minimal additional computation. Our code is available at https://github.com/ cunningham-lab/RTGPS.

****

Dense for the Price of Sparse: Improved Performance of Sparsely Initialized Networks via a Subspace Offset

Ilan Price, Jared Tanner

That neural networks may be pruned to high sparsities and retain high accuracy is well established. Recent research efforts focus on pruning immediately after initialization so as to allow the computational savings afforded by sparsity to extend to the training process. In this work, we introduce a new ‘DCT plus Sparse’ layer architecture, which maintains information propagation and trainability even with as little as 0.01% trainable parameters remaining. We show that standard training of networks built with these layers, and pruned at initialization, achieves state-of-the-art accuracy for extreme sparsities on a variety of benchmark network architectures and datasets. Moreover, these results are achieved using only simple heuristics to determine the locations of the trainable parameters in the network, and thus without having to initially store or compute with the full, unpruned network, as is required by competing prune-at-initialization algorithms. Switching from standard sparse layers to DCT plus Sparse layers does not increase the storage footprint of a network and incurs only a small additional computational overhead.

****

BANG: Bridging Autoregressive and Non-autoregressive Generation with Large Scale Pretraining

Weizhen Qi, Yeyun Gong, Jian Jiao, Yu Yan, Weizhu Chen, Dayiheng Liu, Kewen Tang, Houqiang Li, Jiusheng Chen, Ruofei Zhang, Ming Zhou, Nan Duan

In this paper, we propose BANG, a new pretraining model to Bridge the gap between Autoregressive (AR) and Non-autoregressive (NAR) Generation. AR and NAR generation can be uniformly regarded as to what extent previous tokens can be attended, and BANG bridges AR and NAR generation through designing a novel model structure for large-scale pre-training. A pretrained BANG model can simultaneously support AR, NAR, and semi-NAR generation to meet different requirements. Experiments on question generation (SQuAD 1.1), summarization (XSum), and dialogue generation (PersonaChat) show that BANG improves NAR and semi-NAR performance significantly as well as attaining comparable performance with strong AR pretrained models. Compared with the semi-NAR strong baselines, BANG achieves absolute improvements of 14.01 and 5.24 in the overall scores of SQuAD 1.1 and XSum, respectively. In addition, BANG achieves absolute improvements of 10.73, 6.39, and 5.90 in the overall scores of SQuAD, XSUM, and PersonaChat compared with the NAR strong baselines, respectively. Our code will be made publicly available.

****

A Probabilistic Approach to Neural Network Pruning
Xin Qian, Diego Klabjan
Neural network pruning techniques reduce the number of parameters without compromising predicting ability of a network. Many algorithms have been developed for pruning both over-parameterized fully-connected networks (FCN) and convolutional neural networks (CNN), but analytical studies of capabilities and compression ratios of such pruned sub-networks are lacking. We theoretically study the performance of two pruning techniques (random and magnitude-based) on FCN and CNN. Given a target network, we provide a universal approach to bound the gap between a pruned and the target network in a probabilistic sense, which is the first study of this nature. The results establish that there exist pruned networks with expressive power within any specified bound from the target network and with a significant compression ratio.
****
Global Prosody Style Transfer Without Text Transcriptions
Kaizhi Qian, Yang Zhang, Shiyu Chang, Jinjun Xiong, Chuang Gan, David Cox, Mark Hasegawa-Johnson
Prosody plays an important role in characterizing the style of a speaker or an emotion, but most non-parallel voice or emotion style transfer algorithms do not convert any prosody information. Two major components of prosody are pitch and rhythm. Disentangling the prosody information, particularly the rhythm component, from the speech is challenging because it involves breaking the synchrony between the input speech and the disentangled speech representation. As a result, most existing prosody style transfer algorithms would need to rely on some form of text transcriptions to identify the content information, which confines their application to high-resource languages only. Recently, SpeechSplit has made sizeable progress towards unsupervised prosody style transfer, but it is unable to extract high-level global prosody style in an unsupervised manner. In this paper, we propose AutoPST, which can disentangle global prosody style from speech without relying on any text transcriptions. AutoPST is an Autoencoder-based Prosody Style Transfer framework with a thorough rhythm removal module guided by the self-expressive representation learning. Experiments on different style transfer tasks show that AutoPST can effectively convert prosody that correctly reflects the styles of the target domains.
****
Efficient Differentiable Simulation of Articulated Bodies
Yi-Ling Qiao, Junbang Liang, Vladlen Koltun, Ming C Lin
We present a method for efficient differentiable simulation of articulated bodies. This enables integration of articulated body dynamics into deep learning frameworks, and gradient-based optimization of neural networks that operate on articulated bodies. We derive the gradients of the contact solver using spatial algebra and the adjoint method. Our approach is an order of magnitude faster than autodiff tools. By only saving the initial states throughout the simulation process, our method reduces memory requirements by two orders of magnitude. We demonstrate the utility of efficient differentiable dynamics for articulated bodies in a variety of applications. We show that reinforcement learning with articulated systems can be accelerated using gradients provided by our method. In applications to control and inverse problems, gradient-based optimization enabled by our work accelerates convergence by more than an order of magnitude.
****
Oneshot Differentially Private Top-k Selection
Gang Qiao, Weijie Su, Li Zhang
Being able to efficiently and accurately select the top-k elements with differential privacy is an integral component of various private data analysis tasks. In this paper, we present the oneshot Laplace mechanism, which generalizes the well-known Report Noisy Max \cite{dwork2014algorithmic} mechanism to reporting noisy top-k elements. We show that the oneshot Laplace mechanism with a noise level of $\widetilde{O}(\sqrt{k}/\eps)$ is approximately differentially private. Compared to the previous peeling approach of running Report Noisy Max k times, the oneshot Laplace mechanism only adds noises and computes the top k elements

once, hence much more efficient for large k. In addition, our proof of privacy relies on a novel coupling technique that bypasses the composition theorems so without the linear dependence on k which is inherent to various composition theorems. Finally, we present a novel application of efficient top-k selection in the classical problem of ranking from pairwise comparisons.
****
Density Constrained Reinforcement Learning
Zengyi Qin, Yuxiao Chen, Chuchu Fan
We study constrained reinforcement learning (CRL) from a novel perspective by setting constraints directly on state density functions, rather than the value functions considered by previous works. State density has a clear physical and mathematical interpretation, and is able to express a wide variety of constraints such as resource limits and safety requirements. Density constraints can also avoid the time-consuming process of designing and tuning cost functions required by value function-based constraints to encode system specifications. We leverage the duality between density functions and Q functions to develop an effective algorithm to solve the density constrained RL problem optimally and the constrains are guaranteed to be satisfied. We prove that the proposed algorithm converges to a near-optimal solution with a bounded error even when the policy update is imperfect. We use a set of comprehensive experiments to demonstrate the advantages of our approach over state-of-the-art CRL methods, with a wide range of density constrained tasks as well as standard CRL benchmarks such as Safety-Gym.
****
Budgeted Heterogeneous Treatment Effect Estimation
Tian Qin, Tian-Zuo Wang, Zhi-Hua Zhou
Heterogeneous treatment effect (HTE) estimation is receiving increasing interest due to its important applications in fields such as healthcare, economics, and education. Current HTE estimation methods generally assume the existence of abundant observational data, though the acquisition of such data can be costly. In some real scenarios, it is easy to access the pre-treatment covariates and treatment assignments, but expensive to obtain the factual outcomes. To make HTE estimation more practical, in this paper, we examine the problem of estimating HTEs with a budget constraint on observational data, aiming to obtain accurate HTE estimates with limited costs. By deriving an informative generalization bound and connecting to active learning, we propose an effective and efficient method which is validated both theoretically and empirically.
****
Neural Transformation Learning for Deep Anomaly Detection Beyond Images
Chen Qiu, Timo Pfrommer, Marius Kloft, Stephan Mandt, Maja Rudolph
Data transformations (e.g. rotations, reflections, and cropping) play an important role in self-supervised learning. Typically, images are transformed into different views, and neural networks trained on tasks involving these views produce useful feature representations for downstream tasks, including anomaly detection. However, for anomaly detection beyond image data, it is often unclear which transformations to use. Here we present a simple end-to-end procedure for anomaly detection with learnable transformations. The key idea is to embed the transformed data into a semantic space such that the transformed data still resemble their untransformed form, while different transformations are easily distinguishable. Extensive experiments on time series show that our proposed method outperforms existing approaches in the one-vs.-rest setting and is competitive in the more challenging n-vs.-rest anomaly-detection task. On medical and cyber-security tabular data, our method learns domain-specific transformations and detects anomalies more accurately than previous work.
****
Provably Efficient Fictitious Play Policy Optimization for Zero-Sum Markov Games with Structured Transitions
Shuang Qiu, Xiaohan Wei, Jieping Ye, Zhaoran Wang, Zhuoran Yang
While single-agent policy optimization in a fixed environment has attracted a lot of research attention recently in the reinforcement learning community, much less is known theoretically when there are multiple agents playing in a potential

ly competitive environment. We take steps forward by proposing and analyzing new
 fictitious play policy optimization algorithms for two-player zero-sum Markov g
ames with structured but unknown transitions. We consider two classes of transit
ion structures: factored independent transition and single-controller transition
. For both scenarios, we prove tight $\widetilde{\mathcal{O}}(\sqrt{T})$ regret
bounds after T steps in a two-agent competitive game scenario. The regret of e
ach player is measured against a potentially adversarial opponent who can choose
 a single best policy in hindsight after observing the full policy sequence. Our
 algorithms feature a combination of Upper Confidence Bound (UCB)-type optimism
and fictitious play under the scope of simultaneous policy optimization in a non
-stationary environment. When both players adopt the proposed algorithms, their
overall optimality gap is $\widetilde{\mathcal{O}}(\sqrt{T})$.
****
Optimization Planning for 3D ConvNets
Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Tao Mei
It is not trivial to optimally learn a 3D Convolutional Neural Networks (3D Conv
Nets) due to high complexity and various options of the training scheme. The mos
t common hand-tuning process starts from learning 3D ConvNets using short video
clips and then is followed by learning long-term temporal dependency using lengt
hy clips, while gradually decaying the learning rate from high to low as trainin
g progresses. The fact that such process comes along with several heuristic sett
ings motivates the study to seek an optimal "path" to automate the entire traini
ng. In this paper, we decompose the path into a series of training "states" and
specify the hyper-parameters, e.g., learning rate and the length of input clips,
 in each state. The estimation of the knee point on the performance-epoch curve
triggers the transition from one state to another. We perform dynamic programmin
g over all the candidate states to plan the optimal permutation of states, i.e.,
 optimization path. Furthermore, we devise a new 3D ConvNets with a unique desig
n of dual-head classifier to improve spatial and temporal discrimination. Extens
ive experiments on seven public video recognition benchmarks demonstrate the adv
antages of our proposal. With the optimization planning, our 3D ConvNets achieve
s superior results when comparing to the state-of-the-art recognition methods. M
ore remarkably, we obtain the top-1 accuracy of 80.5% and 82.7% on Kinetics-400
and Kinetics-600 datasets, respectively.
****
On Reward-Free RL with Kernel and Neural Function Approximations: Single-Agent M
DP and Markov Game
Shuang Qiu, Jieping Ye, Zhaoran Wang, Zhuoran Yang
To achieve sample efficiency in reinforcement learning (RL), it necessitates to
efficiently explore the underlying environment. Under the offline setting, addre
ssing the exploration challenge lies in collecting an offline dataset with suffi
cient coverage. Motivated by such a challenge, we study the reward-free RL probl
em, where an agent aims to thoroughly explore the environment without any pre-sp
ecified reward function. Then, given any extrinsic reward, the agent computes th
e optimal policy via offline RL with data collected in the exploration stage. Mo
reover, we tackle this problem under the context of function approximation, leve
raging powerful function approximators. Specifically, we propose to explore via
an optimistic variant of the value-iteration algorithm incorporating kernel and
neural function approximations, where we adopt the associated exploration bonus
as the exploration reward. Moreover, we design exploration and planning algorith
ms for both single-agent MDPs and zero-sum Markov games and prove that our metho
ds can achieve $\widetilde{\mathcal{O}}(1 /\varepsilon^2)$ sample complexity for
 generating a ε-suboptimal policy or ε-approximate Nash
equilibrium when given an arbitrary extrinsic reward. To the best of our knowled
ge, we establish the first provably efficient reward-free RL algorithm with kern
el and neural function approximators.
****
Learning Transferable Visual Models From Natural Language Supervision
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini
 Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Kru

eger, Ilya Sutskever
State-of-the-art computer vision systems are trained to predict a fixed set of p
redetermined object categories. This restricted form of supervision limits their
 generality and usability since additional labeled data is needed to specify any
 other visual concept. Learning directly from raw text about images is a promisi
ng alternative which leverages a much broader source of supervision. We demonstr
ate that the simple pre-training task of predicting which caption goes with whic
h image is an efficient and scalable way to learn SOTA image representations fro
m scratch on a dataset of 400 million (image, text) pairs collected from the int
ernet. After pre-training, natural language is used to reference learned visual
concepts (or describe new ones) enabling zero-shot transfer of the model to down
stream tasks. We study the performance of this approach by benchmarking on over
30 different existing computer vision datasets, spanning tasks such as OCR, acti
on recognition in videos, geo-localization, and many types of fine-grained objec
t classification. The model transfers non-trivially to most tasks and is often c
ompetitive with a fully supervised baseline without the need for any dataset spe
cific training. For instance, we match the accuracy of the original ResNet-50 on
 ImageNet zero-shot without needing to use any of the 1.28 million training exam
ples it was trained on.
****
A General Framework For Detecting Anomalous Inputs to DNN Classifiers
Jayaram Raghuram, Varun Chandrasekaran, Somesh Jha, Suman Banerjee
Detecting anomalous inputs, such as adversarial and out-of-distribution (OOD) in
puts, is critical for classifiers (including deep neural networks or DNNs) deplo
yed in real-world applications. While prior works have proposed various methods
to detect such anomalous samples using information from the internal layer repre
sentations of a DNN, there is a lack of consensus on a principled approach for t
he different components of such a detection method. As a result, often heuristic
 and one-off methods are applied for different aspects of this problem. We propo
se an unsupervised anomaly detection framework based on the internal DNN layer r
epresentations in the form of a meta-algorithm with configurable components. We
proceed to propose specific instantiations for each component of the meta-algori
thm based on ideas grounded in statistical testing and anomaly detection. We eva
luate the proposed methods on well-known image classification datasets with stro
ng adversarial attacks and OOD inputs, including an adaptive attack that uses th
e internal layer representations of the DNN (often not considered in prior work)
. Comparisons with five recently-proposed competing detection methods demonstrat
es the effectiveness of our method in detecting adversarial and OOD inputs.
****
Towards Open Ad Hoc Teamwork Using Graph-based Policy Learning
Muhammad A Rahman, Niklas Hopner, Filippos Christianos, Stefano V Albrecht
Ad hoc teamwork is the challenging problem of designing an autonomous agent whic
h can adapt quickly to collaborate with teammates without prior coordination mec
hanisms, including joint training. Prior work in this area has focused on closed
 teams in which the number of agents is fixed. In this work, we consider open te
ams by allowing agents with different fixed policies to enter and leave the envi
ronment without prior notification. Our solution builds on graph neural networks
 to learn agent models and joint-action value models under varying team composit
ions. We contribute a novel action-value computation that integrates the agent m
odel and joint-action value model to produce action-value estimates. We empirica
lly demonstrate that our approach successfully models the effects other agents h
ave on the learner, leading to policies that robustly adapt to dynamic team comp
ositions and significantly outperform several alternative methods.
****
Decoupling Value and Policy for Generalization in Reinforcement Learning
Roberta Raileanu, Rob Fergus
Standard deep reinforcement learning algorithms use a shared representation for
the policy and value function, especially when training directly from images. Ho
wever, we argue that more information is needed to accurately estimate the value
 function than to learn the optimal policy. Consequently, the use of a shared re

presentation for the policy and value function can lead to overfitting. To alleviate this problem, we propose two approaches which are combined to create IDAAC: Invariant Decoupled Advantage Actor-Critic. First, IDAAC decouples the optimization of the policy and value function, using separate networks to model them. Second, it introduces an auxiliary loss which encourages the representation to be invariant to task-irrelevant properties of the environment. IDAAC shows good generalization to unseen environments, achieving a new state-of-the-art on the Procgen benchmark and outperforming popular methods on DeepMind Control tasks with distractors. Our implementation is available at https://github.com/rraileanu/idaac.
****
Hierarchical Clustering of Data Streams: Scalable Algorithms and Approximation Guarantees
Anand Rajagopalan, Fabio Vitale, Danny Vainstein, Gui Citovsky, Cecilia M Procopiuc, Claudio Gentile
We investigate the problem of hierarchically clustering data streams containing metric data in R^d. We introduce a desirable invariance property for such algorithms, describe a general family of hyperplane-based methods enjoying this property, and analyze two scalable instances of this general family against recently popularized similarity/dissimilarity-based metrics for hierarchical clustering. We prove a number of new results related to the approximation ratios of these algorithms, improving in various ways over the literature on this subject. Finally, since our algorithms are principled but also very practical, we carry out an experimental comparison on both synthetic and real-world datasets showing competitive results against known baselines.
****
Differentially Private Sliced Wasserstein Distance
Alain Rakotomamonjy, Ralaivola Liva
Developing machine learning methods that are privacy preserving is today a central topic of research, with huge practical impacts. Among the numerous ways to address privacy-preserving learning, we here take the perspective of computing the divergences between distributions under the Differential Privacy (DP) framework — being able to compute divergences between distributions is pivotal for many machine learning problems, such as learning generative models or domain adaptation problems. Instead of resorting to the popular gradient-based sanitization method for DP, we tackle the problem at its roots by focusing on the Sliced Wasserstein Distance and seamlessly making it differentially private. Our main contribution is as follows: we analyze the property of adding a Gaussian perturbation to the intrinsic randomized mechanism of the Sliced Wasserstein Distance, and we establish the sensitivity of the resulting differentially private mechanism. One of our important findings is that this DP mechanism transforms the Sliced Wasserstein distance into another distance, that we call the Smoothed Sliced Wasserstein Distance. This new differentially private distribution distance can be plugged into generative models and domain adaptation algorithms in a transparent way, and we empirically show that it yields highly competitive performance compared with gradient-based DP approaches from the literature, with almost no loss in accuracy for the domain adaptation problems that we consider.
****
Zero-Shot Text-to-Image Generation
Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever
Text-to-image generation has traditionally focused on finding better modeling assumptions for training on a fixed dataset. These assumptions might involve complex architectures, auxiliary losses, or side information such as object part labels or segmentation masks supplied during training. We describe a simple approach for this task based on a transformer that autoregressively models the text and image tokens as a single stream of data. With sufficient data and scale, our approach is competitive with previous domain-specific models when evaluated in a zero-shot fashion.
****

End-to-End Learning of Coherent Probabilistic Forecasts for Hierarchical Time Series
Syama Sundar Rangapuram, Lucien D Werner, Konstantinos Benidis, Pedro Mercado, Jan Gasthaus, Tim Januschowski
This paper presents a novel approach for hierarchical time series forecasting that produces coherent, probabilistic forecasts without requiring any explicit post-processing reconciliation. Unlike the state-of-the-art, the proposed method simultaneously learns from all time series in the hierarchy and incorporates the reconciliation step into a single trainable model. This is achieved by applying the reparameterization trick and casting reconciliation as an optimization problem with a closed-form solution. These model features make end-to-end learning of hierarchical forecasts possible, while accomplishing the challenging task of generating forecasts that are both probabilistic and coherent. Importantly, our approach also accommodates general aggregation constraints including grouped and temporal hierarchies. An extensive empirical evaluation on real-world hierarchical datasets demonstrates the advantages of the proposed approach over the state-of-the-art.
****
MSA Transformer
Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, Alexander Rives
Unsupervised protein language models trained across millions of diverse sequences learn structure and function of proteins. Protein language models studied to date have been trained to perform inference from individual sequences. The longstanding approach in computational biology has been to make inferences from a family of evolutionarily related sequences by fitting a model to each family independently. In this work we combine the two paradigms. We introduce a protein language model which takes as input a set of sequences in the form of a multiple sequence alignment. The model interleaves row and column attention across the input sequences and is trained with a variant of the masked language modeling objective across many protein families. The performance of the model surpasses current state-of-the-art unsupervised structure learning methods by a wide margin, with far greater parameter efficiency than prior state-of-the-art protein language models.
****
Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting
Kashif Rasul, Calvin Seward, Ingmar Schuster, Roland Vollgraf
In this work, we propose TimeGrad, an autoregressive model for multivariate probabilistic time series forecasting which samples from the data distribution at each time step by estimating its gradient. To this end, we use diffusion probabilistic models, a class of latent variable models closely connected to score matching and energy-based methods. Our model learns gradients by optimizing a variational bound on the data likelihood and at inference time converts white noise into a sample of the distribution of interest through a Markov chain using Langevin sampling. We demonstrate experimentally that the proposed autoregressive denoising diffusion model is the new state-of-the-art multivariate probabilistic forecasting method on real-world data sets with thousands of correlated dimensions. We hope that this method is a useful tool for practitioners and lays the foundation for future research in this area.
****
Generative Particle Variational Inference via Estimation of Functional Gradients
Neale Ratzlaff, Qinxun Bai, Li Fuxin, Wei Xu
Recently, particle-based variational inference (ParVI) methods have gained interest because they can avoid arbitrary parametric assumptions that are common in variational inference. However, many ParVI approaches do not allow arbitrary sampling from the posterior, and the few that do allow such sampling suffer from suboptimality. This work proposes a new method for learning to approximately sample from the posterior distribution. We construct a neural sampler that is trained with the functional gradient of the KL-divergence between the empirical sampling

distribution and the target distribution, assuming the gradient resides within a reproducing kernel Hilbert space. Our generative ParVI (GPVI) approach maintains the asymptotic performance of ParVI methods while offering the flexibility of a generative sampler. Through carefully constructed experiments, we show that GPVI outperforms previous generative ParVI methods such as amortized SVGD, and is competitive with ParVI as well as gold-standard approaches like Hamiltonian Monte Carlo for fitting both exactly known and intractable target distributions.
****
Enhancing Robustness of Neural Networks through Fourier Stabilization
Netanel Raviv, Aidan Kelley, Minzhe Guo, Yevgeniy Vorobeychik
Despite the considerable success of neural networks in security settings such as malware detection, such models have proved vulnerable to evasion attacks, in which attackers make slight changes to inputs (e.g., malware) to bypass detection. We propose a novel approach, Fourier stabilization, for designing evasion-robust neural networks with binary inputs. This approach, which is complementary to other forms of defense, replaces the weights of individual neurons with robust analogs derived using Fourier analytic tools. The choice of which neurons to stabilize in a neural network is then a combinatorial optimization problem, and we propose several methods for approximately solving it. We provide a formal bound on the per-neuron drop in accuracy due to Fourier stabilization, and experimentally demonstrate the effectiveness of the proposed approach in boosting robustness of neural networks in several detection settings. Moreover, we show that our approach effectively composes with adversarial training.
****
Disentangling Sampling and Labeling Bias for Learning in Large-output Spaces
Ankit Singh Rawat, Aditya K Menon, Wittawat Jitkrittum, Sadeep Jayasumana, Felix Yu, Sashank Reddi, Sanjiv Kumar
Negative sampling schemes enable efficient training given a large number of classes, by offering a means to approximate a computationally expensive loss function that takes all labels into account. In this paper, we present a new connection between these schemes and loss modification techniques for countering label imbalance. We show that different negative sampling schemes implicitly trade-off performance on dominant versus rare labels. Further, we provide a unified means to explicitly tackle both sampling bias, arising from working with a subset of all labels, and labeling bias, which is inherent to the data due to label imbalance. We empirically verify our findings on long-tail classification and retrieval benchmarks.
****
Cross-domain Imitation from Observations
Dripta S. Raychaudhuri, Sujoy Paul, Jeroen Vanbaar, Amit K. Roy-Chowdhury
Imitation learning seeks to circumvent the difficulty in designing proper reward functions for training agents by utilizing expert behavior. With environments modeled as Markov Decision Processes (MDP), most of the existing imitation algorithms are contingent on the availability of expert demonstrations in the same MDP as the one in which a new imitation policy is to be learned. In this paper, we study the problem of how to imitate tasks when discrepancies exist between the expert and agent MDP. These discrepancies across domains could include differing dynamics, viewpoint, or morphology; we present a novel framework to learn correspondences across such domains. Importantly, in contrast to prior works, we use unpaired and unaligned trajectories containing only states in the expert domain, to learn this correspondence. We utilize a cycle-consistency constraint on both the state space and a domain agnostic latent space to do this. In addition, we enforce consistency on the temporal position of states via a normalized position estimator function, to align the trajectories across the two domains. Once this correspondence is found, we can directly transfer the demonstrations on one domain to the other and use it for imitation. Experiments across a wide variety of challenging domains demonstrate the efficacy of our approach.
****
Implicit Regularization in Tensor Factorization
Noam Razin, Asaf Maman, Nadav Cohen

Recent efforts to unravel the mystery of implicit regularization in deep learnin
g have led to a theoretical focus on matrix factorization — matrix completion vi
a linear neural network. As a step further towards practical deep learning, we p
rovide the first theoretical analysis of implicit regularization in tensor facto
rization — tensor completion via certain type of non-linear neural network. We c
ircumvent the notorious difficulty of tensor problems by adopting a dynamical sy
stems perspective, and characterizing the evolution induced by gradient descent.
 The characterization suggests a form of greedy low tensor rank search, which we
 rigorously prove under certain conditions, and empirically demonstrate under ot
hers. Motivated by tensor rank capturing the implicit regularization of a non-li
near neural network, we empirically explore it as a measure of complexity, and f
ind that it captures the essence of datasets on which neural networks generalize
. This leads us to believe that tensor rank may pave way to explaining both impl
icit regularization in deep learning, and the properties of real-world data tran
slating this implicit regularization to generalization.
****
Align, then memorise: the dynamics of learning with feedback alignment
Maria Refinetti, Stéphane D’Ascoli, Ruben Ohana, Sebastian Goldt
Direct Feedback Alignment (DFA) is emerging as an efficient and biologically pla
usible alternative to backpropagation for training deep neural networks. Despite
 relying on random feedback weights for the backward pass, DFA successfully trai
ns state-of-the-art models such as Transformers. On the other hand, it notorious
ly fails to train convolutional networks. An understanding of the inner workings
 of DFA to explain these diverging results remains elusive. Here, we propose a t
heory of feedback alignment algorithms. We first show that learning in shallow n
etworks proceeds in two steps: an alignment phase, where the model adapts its we
ights to align the approximate gradient with the true gradient of the loss funct
ion, is followed by a memorisation phase, where the model focuses on fitting the
 data. This two-step process has a degeneracy breaking effect: out of all the lo
w-loss solutions in the landscape, a net-work trained with DFA naturally converg
es to the solution which maximises gradient alignment. We also identify a key qu
antity underlying alignment in deep linear networks: the conditioning of the ali
gnment matrices. The latter enables a detailed understanding of the impact of da
ta structure on alignment, and suggests a simple explanation for the well-known
failure of DFA to train convolutional neural networks. Numerical experiments on
MNIST and CIFAR10 clearly demonstrate degeneracy breaking in deep non-linear net
works and show that the align-then-memorize process occurs sequentially from the
 bottom layers of the network to the top.
****
Classifying high-dimensional Gaussian mixtures: Where kernel methods fail and ne
ural networks succeed
Maria Refinetti, Sebastian Goldt, Florent Krzakala, Lenka Zdeborova
A recent series of theoretical works showed that the dynamics of neural networks
 with a certain initialisation are well-captured by kernel methods. Concurrent e
mpirical work demonstrated that kernel methods can come close to the performance
 of neural networks on some image classification tasks. These results raise the
question of whether neural networks only learn successfully if kernels also lear
n successfully, despite being the more expressive function class. Here, we show
that two-layer neural networks with *only a few neurons* achieve near-optimal pe
rformance on high-dimensional Gaussian mixture classification while lazy trainin
g approaches such as random features and kernel methods do not. Our analysis is
based on the derivation of a set of ordinary differential equations that exactly
 track the dynamics of the network and thus allow to extract the asymptotic perf
ormance of the network as a function of regularisation or signal-to-noise ratio.
 We also show how over-parametrising the neural network leads to faster converge
nce, but does not improve its final performance.
****
Sharf: Shape-conditioned Radiance Fields from a Single View
Konstantinos Rematas, Ricardo Martin-Brualla, Vittorio Ferrari
We present a method for estimating neural scenes representations of objects give

n only a single image. The core of our method is the estimation of a geometric s caffold for the object and its use as a guide for the reconstruction of the unde rlying radiance field. Our formulation is based on a generative process that fir st maps a latent code to a voxelized shape, and then renders it to an image, wit h the object appearance being controlled by a second latent code. During inferen ce, we optimize both the latent codes and the networks to fit a test image of a new object. The explicit disentanglement of shape and appearance allows our mode l to be fine-tuned given a single image. We can then render new views in a geome trically consistent manner and they represent faithfully the input object. Addit ionally, our method is able to generalize to images outside of the training doma in (more realistic renderings and even real photographs). Finally, the inferred geometric scaffold is itself an accurate estimate of the object’s 3D shape. We d emonstrate in several experiments the effectiveness of our approach in both synt hetic and real images.
****
LEGO: Latent Execution-Guided Reasoning for Multi-Hop Question Answering on Know ledge Graphs
Hongyu Ren, Hanjun Dai, Bo Dai, Xinyun Chen, Michihiro Yasunaga, Haitian Sun, Da le Schuurmans, Jure Leskovec, Denny Zhou
Answering complex natural language questions on knowledge graphs (KGQA) is a cha llenging task. It requires reasoning with the input natural language questions a s well as a massive, incomplete heterogeneous KG. Prior methods obtain an abstra ct structured query graph/tree from the input question and traverse the KG for a nswers following the query tree. However, they inherently cannot deal with missi ng links in the KG. Here we present LEGO, a Latent Execution-Guided reasOning fr amework to handle this challenge in KGQA. LEGO works in an iterative way, which alternates between (1) a Query Synthesizer, which synthesizes a reasoning action and grows the query tree step-by-step, and (2) a Latent Space Executor that exe cutes the reasoning action in the latent embedding space to combat against the m issing information in KG. To learn the synthesizer without step-wise supervision , we design a generic latent execution guided bottom-up search procedure to find good execution traces efficiently in the vast query space. Experimental results on several KGQA benchmarks demonstrate the effectiveness of our framework compa red with previous state of the art.
****
Interpreting and Disentangling Feature Components of Various Complexity from DNN s
Jie Ren, Mingjie Li, Zexu Liu, Quanshi Zhang
This paper aims to define, visualize, and analyze the feature complexity that is learned by a DNN. We propose a generic definition for the feature complexity. G iven the feature of a certain layer in the DNN, our method decomposes and visual izes feature components of different complexity orders from the feature. The fea ture decomposition enables us to evaluate the reliability, the effectiveness, an d the significance of over-fitting of these feature components. Furthermore, suc h analysis helps to improve the performance of DNNs. As a generic method, the fe ature complexity also provides new insights into existing deep-learning techniqu es, such as network compression and knowledge distillation.
****
Integrated Defense for Resilient Graph Matching
Jiaxiang Ren, Zijie Zhang, Jiayin Jin, Xin Zhao, Sixing Wu, Yang Zhou, Yelong Sh en, Tianshi Che, Ruoming Jin, Dejing Dou
A recent study has shown that graph matching models are vulnerable to adversaria l manipulation of their input which is intended to cause a mismatching. Neverthe less, there is still a lack of a comprehensive solution for further enhancing th e robustness of graph matching against adversarial attacks. In this paper, we id entify and study two types of unique topology attacks in graph matching: inter-g raph dispersion and intra-graph assembly attacks. We propose an integrated defen se model, IDRGM, for resilient graph matching with two novel defense techniques to defend against the above two attacks simultaneously. A detection technique of inscribed simplexes in the hyperspheres consisting of multiple matched nodes is

proposed to tackle inter-graph dispersion attacks, in which the distances among the matched nodes in multiple graphs are maximized to form regular simplexes. A node separation method based on phase-type distribution and maximum likelihood estimation is developed to estimate the distribution of perturbed graphs and separate the nodes within the same graphs over a wide space, for defending intra-graph assembly attacks, such that the interference from the similar neighbors of the perturbed nodes is significantly reduced. We evaluate the robustness of our IDRGM model on real datasets against state-of-the-art algorithms.
****
Solving high-dimensional parabolic PDEs using the tensor train format
Lorenz Richter, Leon Sallandt, Nikolas Nüsken
High-dimensional partial differential equations (PDEs) are ubiquitous in economics, science and engineering. However, their numerical treatment poses formidable challenges since traditional grid-based methods tend to be frustrated by the curse of dimensionality. In this paper, we argue that tensor trains provide an appealing approximation framework for parabolic PDEs: the combination of reformulations in terms of backward stochastic differential equations and regression-type methods in the tensor format holds the promise of leveraging latent low-rank structures enabling both compression and efficient computation. Following this paradigm, we develop novel iterative schemes, involving either explicit and fast or implicit and accurate updates. We demonstrate in a number of examples that our methods achieve a favorable trade-off between accuracy and computational efficiency in comparison with state-of-the-art neural network based approaches.
****
Best Arm Identification in Graphical Bilinear Bandits
Geovani Rizk, Albert Thomas, Igor Colin, Rida Laraki, Yann Chevaleyre
We introduce a new graphical bilinear bandit problem where a learner (or a \emph{central entity}) allocates arms to the nodes of a graph and observes for each edge a noisy bilinear reward representing the interaction between the two end nodes. We study the best arm identification problem in which the learner wants to find the graph allocation maximizing the sum of the bilinear rewards. By efficiently exploiting the geometry of this bandit problem, we propose a \emph{decentralized} allocation strategy based on random sampling with theoretical guarantees. In particular, we characterize the influence of the graph structure (e.g. star, complete or circle) on the convergence rate and propose empirical experiments that confirm this dependency.
****
Principled Simplicial Neural Networks for Trajectory Prediction
T. Mitchell Roddenberry, Nicholas Glaze, Santiago Segarra
We consider the construction of neural network architectures for data on simplicial complexes. In studying maps on the chain complex of a simplicial complex, we define three desirable properties of a simplicial neural network architecture: namely, permutation equivariance, orientation equivariance, and simplicial awareness. The first two properties respectively account for the fact that the node indexing and the simplex orientations in a simplicial complex are arbitrary. The last property encodes the desirable feature that the output of the neural network depends on the entire simplicial complex and not on a subset of its dimensions. Based on these properties, we propose a simple convolutional architecture, rooted in tools from algebraic topology, for the problem of trajectory prediction, and show that it obeys all three of these properties when an odd, nonlinear activation function is used. We then demonstrate the effectiveness of this architecture in extrapolating trajectories on synthetic and real datasets, with particular emphasis on the gains in generalizability to unseen trajectories.
****
On Linear Identifiability of Learned Representations
Geoffrey Roeder, Luke Metz, Durk Kingma
Identifiability is a desirable property of a statistical model: it implies that the true model parameters may be estimated to any desired precision, given sufficient computational resources and data. We study identifiability in the context of representation learning: discovering nonlinear data representations that are

optimal with respect to some downstream task. When parameterized as deep neural networks, such representation functions lack identifiability in parameter space, because they are over-parameterized by design. In this paper, building on recent advances in nonlinear Independent Components Analysis, we aim to rehabilitate identifiability by showing that a large family of discriminative models are in fact identifiable in function space, up to a linear indeterminacy. Many models for representation learning in a wide variety of domains have been identifiable in this sense, including text, images and audio, state-of-the-art at time of publication. We derive sufficient conditions for linear identifiability and provide empirical support for the result on both simulated and real-world data.
****
Representation Matters: Assessing the Importance of Subgroup Allocations in Training Data
Esther Rolf, Theodora T Worledge, Benjamin Recht, Michael Jordan
Collecting more diverse and representative training data is often touted as a remedy for the disparate performance of machine learning predictors across subpopulations. However, a precise framework for understanding how dataset properties like diversity affect learning outcomes is largely lacking. By casting data collection as part of the learning process, we demonstrate that diverse representation in training data is key not only to increasing subgroup performances, but also to achieving population-level objectives. Our analysis and experiments describe how dataset compositions influence performance and provide constructive results for using trends in existing data, alongside domain knowledge, to help guide intentional, objective-aware dataset design
****
TeachMyAgent: a Benchmark for Automatic Curriculum Learning in Deep RL
Clément Romac, Rémy Portelas, Katja Hofmann, Pierre-Yves Oudeyer
Training autonomous agents able to generalize to multiple tasks is a key target of Deep Reinforcement Learning (DRL) research. In parallel to improving DRL algorithms themselves, Automatic Curriculum Learning (ACL) study how teacher algorithms can train DRL agents more efficiently by adapting task selection to their evolving abilities. While multiple standard benchmarks exist to compare DRL agents, there is currently no such thing for ACL algorithms. Thus, comparing existing approaches is difficult, as too many experimental parameters differ from paper to paper. In this work, we identify several key challenges faced by ACL algorithms. Based on these, we present TeachMyAgent (TA), a benchmark of current ACL algorithms leveraging procedural task generation. It includes 1) challenge-specific unit-tests using variants of a procedural Box2D bipedal walker environment, and 2) a new procedural Parkour environment combining most ACL challenges, making it ideal for global performance assessment. We then use TeachMyAgent to conduct a comparative study of representative existing approaches, showcasing the competitiveness of some ACL algorithms that do not use expert knowledge. We also show that the Parkour environment remains an open problem. We open-source our environments, all studied ACL algorithms (collected from open-source code or re-implemented), and DRL students in a Python package available at https://github.com/flowersteam/TeachMyAgent.
****
Discretization Drift in Two-Player Games
Mihaela C Rosca, Yan Wu, Benoit Dherin, David Barrett
Gradient-based methods for two-player games produce rich dynamics that can solve challenging problems, yet can be difficult to stabilize and understand. Part of this complexity originates from the discrete update steps given by simultaneous or alternating gradient descent, which causes each player to drift away from the continuous gradient flow – a phenomenon we call discretization drift. Using backward error analysis, we derive modified continuous dynamical systems that closely follow the discrete dynamics. These modified dynamics provide an insight into the notorious challenges associated with zero-sum games, including Generative Adversarial Networks. In particular, we identify distinct components of the discretization drift that can alter performance and in some cases destabilize the game. Finally, quantifying discretization drift allows us to identify regularizers

that explicitly cancel harmful forms of drift or strengthen beneficial forms of drift, and thus improve performance of GAN training.
****
On the Predictability of Pruning Across Scales
Jonathan S Rosenfeld, Jonathan Frankle, Michael Carbin, Nir Shavit
We show that the error of iteratively magnitude-pruned networks empirically follows a scaling law with interpretable coefficients that depend on the architecture and task. We functionally approximate the error of the pruned networks, showing it is predictable in terms of an invariant tying width, depth, and pruning level, such that networks of vastly different pruned densities are interchangeable. We demonstrate the accuracy of this approximation over orders of magnitude in depth, width, dataset size, and density. We show that the functional form holds (generalizes) for large scale data (e.g., ImageNet) and architectures (e.g., ResNets). As neural networks become ever larger and costlier to train, our findings suggest a framework for reasoning conceptually and analytically about a standard method for unstructured pruning.
****
Benchmarks, Algorithms, and Metrics for Hierarchical Disentanglement
Andrew Ross, Finale Doshi-Velez
In representation learning, there has been recent interest in developing algorithms to disentangle the ground-truth generative factors behind a dataset, and metrics to quantify how fully this occurs. However, these algorithms and metrics often assume that both representations and ground-truth factors are flat, continuous, and factorized, whereas many real-world generative processes involve rich hierarchical structure, mixtures of discrete and continuous variables with dependence between them, and even varying intrinsic dimensionality. In this work, we develop benchmarks, algorithms, and metrics for learning such hierarchical representations.
****
Simultaneous Similarity-based Self-Distillation for Deep Metric Learning
Karsten Roth, Timo Milbich, Bjorn Ommer, Joseph Paul Cohen, Marzyeh Ghassemi
Deep Metric Learning (DML) provides a crucial tool for visual similarity and zero-shot retrieval applications by learning generalizing embedding spaces, although recent work in DML has shown strong performance saturation across training objectives. However, generalization capacity is known to scale with the embedding space dimensionality. Unfortunately, high dimensional embeddings also create higher retrieval cost for downstream applications. To remedy this, we propose S2SD - Simultaneous Similarity-based Self-distillation. S2SD extends DML with knowledge distillation from auxiliary, high-dimensional embedding and feature spaces to leverage complementary context during training while retaining test-time cost and with negligible changes to the training time. Experiments and ablations across different objectives and standard benchmarks show S2SD offering highly significant improvements of up to 7% in Recall@1, while also setting a new state-of-the-art.
****
Multi-group Agnostic PAC Learnability
Guy N Rothblum, Gal Yona
An agnostic PAC learning algorithm finds a predictor that is competitive with the best predictor in a benchmark hypothesis class, where competitiveness is measured with respect to a given loss function. However, its predictions might be quite sub-optimal for structured subgroups of individuals, such as protected demographic groups. Motivated by such fairness concerns, we study “multi-group agnostic PAC learnability”: fixing a measure of loss, a benchmark class $\H$ and a (potentially) rich collection of subgroups $\G$, the objective is to learn a single predictor such that the loss experienced by every group $g \in \G$ is not much larger than the best possible loss for this group within $\H$. Under natural conditions, we provide a characterization of the loss functions for which such a predictor is guaranteed to exist. For any such loss function we construct a learning algorithm whose sample complexity is logarithmic in the size of the collection $\G$. Our results unify and extend previous positive and negative results from

the multi-group fairness literature, which applied for specific loss functions.
****

PACOH: Bayes-Optimal Meta-Learning with PAC-Guarantees
Jonas Rothfuss, Vincent Fortuin, Martin Josifoski, Andreas Krause
Meta-learning can successfully acquire useful inductive biases from data. Yet, its generalization properties to unseen learning tasks are poorly understood. Particularly if the number of meta-training tasks is small, this raises concerns about overfitting. We provide a theoretical analysis using the PAC-Bayesian framework and derive novel generalization bounds for meta-learning. Using these bounds, we develop a class of PAC-optimal meta-learning algorithms with performance guarantees and a principled meta-level regularization. Unlike previous PAC-Bayesian meta-learners, our method results in a standard stochastic optimization problem which can be solved efficiently and scales well.When instantiating our PAC-optimal hyper-posterior (PACOH) with Gaussian processes and Bayesian Neural Networks as base learners, the resulting methods yield state-of-the-art performance, both in terms of predictive accuracy and the quality of uncertainty estimates. Thanks to their principled treatment of uncertainty, our meta-learners can also be successfully employed for sequential decision problems.
****

An Algorithm for Stochastic and Adversarial Bandits with Switching Costs
Chloé Rouyer, Yevgeny Seldin, Nicolò Cesa-Bianchi
We propose an algorithm for stochastic and adversarial multiarmed bandits with switching costs, where the algorithm pays a price λ every time it switches the arm being played. Our algorithm is based on adaptation of the Tsallis-INF algorithm of Zimmert and Seldin (2021) and requires no prior knowledge of the regime or time horizon. In the oblivious adversarial setting it achieves the minimax optimal regret bound of $ O((\lambda K)^{1/3}T^{2/3} + \sqrt{KT})$, where T is the time horizon and K is the number of arms. In the stochastically constrained adversarial regime, which includes the stochastic regime as a special case, it achieves a regret bound of $O((\lambda K)^{2/3} T^{1/3} + \ln T)\sum_{i \neq i^*} \Delta_i^{-1})$, where Δ_i are suboptimality gaps and i^* is the unique optimal arm. In the special case of $\lambda = 0$ (no switching costs), both bounds are minimax optimal within constants. We also explore variants of the problem, where switching cost is allowed to change over time. We provide experimental evaluation showing competitiveness of our algorithm with the relevant baselines in the stochastic, stochastically constrained adversarial, and adversarial regimes with fixed switching cost.
****

Improving Lossless Compression Rates via Monte Carlo Bits-Back Coding
Yangjun Ruan, Karen Ullrich, Daniel S Severo, James Townsend, Ashish Khisti, Arnaud Doucet, Alireza Makhzani, Chris Maddison
Latent variable models have been successfully applied in lossless compression with the bits-back coding algorithm. However, bits-back suffers from an increase in the bitrate equal to the KL divergence between the approximate posterior and the true posterior. In this paper, we show how to remove this gap asymptotically by deriving bits-back coding algorithms from tighter variational bounds. The key idea is to exploit extended space representations of Monte Carlo estimators of the marginal likelihood. Naively applied, our schemes would require more initial bits than the standard bits-back coder, but we show how to drastically reduce this additional cost with couplings in the latent space. When parallel architectures can be exploited, our coders can achieve better rates than bits-back with little additional cost. We demonstrate improved lossless compression rates in a variety of settings, especially in out-of-distribution or sequential data compression.
****

On Signal-to-Noise Ratio Issues in Variational Inference for Deep Gaussian Processes
Tim G. J. Rudner, Oscar Key, Yarin Gal, Tom Rainforth
We show that the gradient estimates used in training Deep Gaussian Processes (DGPs) with importance-weighted variational inference are susceptible to signal-to-

noise ratio (SNR) issues. Specifically, we show both theoretically and via an extensive empirical evaluation that the SNR of the gradient estimates for the latent variable's variational parameters decreases as the number of importance samples increases. As a result, these gradient estimates degrade to pure noise if the number of importance samples is too large. To address this pathology, we show how doubly-reparameterized gradient estimators, originally proposed for training variational autoencoders, can be adapted to the DGP setting and that the resultant estimators completely remedy the SNR issue, thereby providing more reliable training. Finally, we demonstrate that our fix can lead to consistent improvements in the predictive performance of DGP models.
****
Tilting the playing field: Dynamical loss functions for machine learning
Miguel Ruiz-Garcia, Ge Zhang, Samuel S Schoenholz, Andrea J. Liu
We show that learning can be improved by using loss functions that evolve cyclically during training to emphasize one class at a time. In underparameterized networks, such dynamical loss functions can lead to successful training for networks that fail to find deep minima of the standard cross-entropy loss. In overparameterized networks, dynamical loss functions can lead to better generalization. Improvement arises from the interplay of the changing loss landscape with the dynamics of the system as it evolves to minimize the loss. In particular, as the loss function oscillates, instabilities develop in the form of bifurcation cascades, which we study using the Hessian and Neural Tangent Kernel. Valleys in the landscape widen and deepen, and then narrow and rise as the loss landscape changes during a cycle. As the landscape narrows, the learning rate becomes too large and the network becomes unstable and bounces around the valley. This process ultimately pushes the system into deeper and wider regions of the loss landscape and is characterized by decreasing eigenvalues of the Hessian. This results in better regularized models with improved generalization performance.
****
UnICORNN: A recurrent model for learning very long time dependencies
T. Konstantin Rusch, Siddhartha Mishra
The design of recurrent neural networks (RNNs) to accurately process sequential inputs with long-time dependencies is very challenging on account of the exploding and vanishing gradient problem. To overcome this, we propose a novel RNN architecture which is based on a structure preserving discretization of a Hamiltonian system of second-order ordinary differential equations that models networks of oscillators. The resulting RNN is fast, invertible (in time), memory efficient and we derive rigorous bounds on the hidden state gradients to prove the mitigation of the exploding and vanishing gradient problem. A suite of experiments are presented to demonstrate that the proposed RNN provides state of the art performance on a variety of learning tasks with (very) long-time dependencies.
****
Simple and Effective VAE Training with Calibrated Decoders
Oleh Rybkin, Kostas Daniilidis, Sergey Levine
Variational autoencoders (VAEs) provide an effective and simple method for modeling complex distributions. However, training VAEs often requires considerable hyperparameter tuning to determine the optimal amount of information retained by the latent variable. We study the impact of calibrated decoders, which learn the uncertainty of the decoding distribution and can determine this amount of information automatically, on the VAE performance. While many methods for learning calibrated decoders have been proposed, many of the recent papers that employ VAEs rely on heuristic hyperparameters and ad-hoc modifications instead. We perform the first comprehensive comparative analysis of calibrated decoder and provide recommendations for simple and effective VAE training. Our analysis covers a range of datasets and several single-image and sequential VAE models. We further propose a simple but novel modification to the commonly used Gaussian decoder, which computes the prediction variance analytically. We observe empirically that using heuristic modifications is not necessary with our method.
****
Model-Based Reinforcement Learning via Latent-Space Collocation

Oleh Rybkin, Chuning Zhu, Anusha Nagabandi, Kostas Daniilidis, Igor Mordatch, Se
rgey Levine
The ability to plan into the future while utilizing only raw high-dimensional observations, such as images, can provide autonomous agents with broad and general capabilities. However, realistic tasks require performing temporally extended reasoning, and cannot be solved with only myopic, short-sighted planning. Recent work in model-based reinforcement learning (RL) has shown impressive results on tasks that require only short-horizon reasoning. In this work, we study how the long-horizon planning abilities can be improved with an algorithm that optimizes over sequences of states, rather than actions, which allows better credit assignment. To achieve this, we draw on the idea of collocation and adapt it to the image-based setting by leveraging probabilistic latent variable models, resulting in an algorithm that optimizes trajectories over latent variables. Our latent collocation method (LatCo) provides a general and effective visual planning approach, and significantly outperforms prior model-based approaches on challenging visual control tasks with sparse rewards and long-term goals. See the videos on the supplementary website \url{https://sites.google.com/view/latco-mbrl/.}
****
Training Data Subset Selection for Regression with Controlled Generalization Error
Durga S, Rishabh Iyer, Ganesh Ramakrishnan, Abir De
Data subset selection from a large number of training instances has been a successful approach toward efficient and cost-effective machine learning. However, models trained on a smaller subset may show poor generalization ability. In this paper, our goal is to design an algorithm for selecting a subset of the training data, so that the model can be trained quickly, without significantly sacrificing on accuracy. More specifically, we focus on data subset selection for L_2 regularized regression problems and provide a novel problem formulation which seeks to minimize the training loss with respect to both the trainable parameters and the subset of training data, subject to error bounds on the validation set. We tackle this problem using several technical innovations. First, we represent this problem with simplified constraints using the dual of the original training problem and show that the objective of this new representation is a monotone and α-submodular function, for a wide variety of modeling choices. Such properties lead us to develop SELCON, an efficient majorization-minimization algorithm for data subset selection, that admits an approximation guarantee even when the training provides an imperfect estimate of the trained model. Finally, our experiments on several datasets show that SELCON trades off accuracy and efficiency more effectively than the current state-of-the-art.
****
Unsupervised Part Representation by Flow Capsules
Sara Sabour, Andrea Tagliasacchi, Soroosh Yazdani, Geoffrey Hinton, David J Fleet
Capsule networks aim to parse images into a hierarchy of objects, parts and relations. While promising, they remain limited by an inability to learn effective low level part descriptions. To address this issue we propose a way to learn primary capsule encoders that detect atomic parts from a single image. During training we exploit motion as a powerful perceptual cue for part definition, with an expressive decoder for part generation within a layered image model with occlusion. Experiments demonstrate robust part discovery in the presence of multiple objects, cluttered backgrounds, and occlusion. The learned part decoder is shown to infer the underlying shape masks, effectively filling in occluded regions of the detected shapes. We evaluate FlowCapsules on unsupervised part segmentation and unsupervised image classification.
****
Stochastic Sign Descent Methods: New Algorithms and Better Theory
Mher Safaryan, Peter Richtarik
Various gradient compression schemes have been proposed to mitigate the communication cost in distributed training of large scale machine learning models. Sign-based methods, such as signSGD (Bernstein et al., 2018), have recently been gain

ing popularity because of their simple compression rule and connection to adaptive gradient methods, like ADAM. In this paper, we analyze sign-based methods for non-convex optimization in three key settings: (i) standard single node, (ii) parallel with shared data and (iii) distributed with partitioned data. For single machine case, we generalize the previous analysis of signSGD relying on intuitive bounds on success probabilities and allowing even biased estimators. Furthermore, we extend the analysis to parallel setting within a parameter server framework, where exponentially fast noise reduction is guaranteed with respect to number of nodes, maintaining 1-bit compression in both directions and using small mini-batch sizes. Next, we identify a fundamental issue with signSGD to converge in distributed environment. To resolve this issue, we propose a new sign-based method, {\em Stochastic Sign Descent with Momentum (SSDM)}, which converges under standard bounded variance assumption with the optimal asymptotic rate. We validate several aspects of our theoretical findings with numerical experiments.
****
Adversarial Dueling Bandits
Aadirupa Saha, Tomer Koren, Yishay Mansour
We introduce the problem of regret minimization in Adversarial Dueling Bandits. As in classic Dueling Bandits, the learner has to repeatedly choose a pair of items and observe only a relative binary ‘win-loss’ feedback for this pair, but here this feedback is generated from an arbitrary preference matrix, possibly chosen adversarially. Our main result is an algorithm whose T-round regret compared to the \emph{Borda-winner} from a set of K items is $\tilde{O}(K^{1/3}T^{2/3})$, as well as a matching $\Omega(K^{1/3}T^{2/3})$ lower bound. We also prove a similar high probability regret bound. We further consider a simpler \emph{fixed-gap} adversarial setup, which bridges between two extreme preference feedback models for dueling bandits: stationary preferences and an arbitrary sequence of preferences. For the fixed-gap adversarial setup we give an $\smash{ \tilde{O}((K/\Delta^2)\log{T}) }$ regret algorithm, where Δ is the gap in Borda scores between the best item and all other items, and show a lower bound of $\Omega(K/\Delta^2)$ indicating that our dependence on the main problem parameters K and Δ is tight (up to logarithmic factors). Finally, we corroborate the theoretical results with empirical evaluations.
****
Dueling Convex Optimization
Aadirupa Saha, Tomer Koren, Yishay Mansour
We address the problem of convex optimization with preference (dueling) feedback. Like the traditional optimization objective, the goal is to find the optimal point with the least possible query complexity, however, without the luxury of even a zeroth order feedback. Instead, the learner can only observe a single noisy bit which is win-loss feedback for a pair of queried points based on their function values. % The problem is certainly of great practical relevance as in many real-world scenarios, such as recommender systems or learning from customer preferences, where the system feedback is often restricted to just one binary-bit preference information. % We consider the problem of online convex optimization (OCO) solely by actively querying $\{0,1\}$ noisy-comparison feedback of decision point pairs, with the objective of finding a near-optimal point (function minimizer) with the least possible number of queries. %a very general class of monotonic, non-decreasing transfer functions, and analyze the problem for any d-dimensional smooth convex function. % For the non-stationary OCO setup, where the underlying convex function may change over time, we prove an impossibility result towards achieving the above objective. We next focus only on the stationary OCO problem, and our main contribution lies in designing a normalized gradient descent based algorithm towards finding a ϵ-best optimal point. Towards this, our algorithm is shown to yield a convergence rate of $\tilde O(\nicefrac{d\beta}{\epsilon \nu^2})$ (ν being the noise parameter) when the underlying function is β-smooth. Further we show an improved convergence rate of just $\tilde O(\nicefrac{d\beta}{\alpha \nu^2} \log \frac{1}{\epsilon})$ when the function is additionally also α-strongly convex.
****

Optimal regret algorithm for Pseudo-1d Bandit Convex Optimization
Aadirupa Saha, Nagarajan Natarajan, Praneeth Netrapalli, Prateek Jain
We study online learning with bandit feedback (i.e. learner has access to only zeroth-order oracle) where cost/reward functions $\f_t$ admit a "pseudo-1d" structure, i.e. $\f_t(\w) = \loss_t(\pred_t(\w))$ where the output of $\pred_t$ is one-dimensional. At each round, the learner observes context $\x_t$, plays prediction $\pred_t(\w_t; \x_t)$ (e.g. $\pred_t(\cdot)=■\x_t, \cdot■$) for some $\w_t \in \mathbb{R}^d$ and observes loss $\loss_t(\pred_t(\w_t))$ where $\loss_t$ is a convex Lipschitz-continuous function. The goal is to minimize the standard regret metric. This pseudo-1d bandit convex optimization problem (\SBCO) arises frequently in domains such as online decision-making or parameter-tuning in large systems. For this problem, we first show a regret lower bound of $\min(\sqrt{dT}, T^{3/4})$ for any algorithm, where T is the number of rounds. We propose a new algorithm \sbcalg that combines randomized online gradient descent with a kernelized exponential weights method to exploit the pseudo-1d structure effectively, guaranteeing the {\em optimal} regret bound mentioned above, up to additional logarithmic factors. In contrast, applying state-of-the-art online convex optimization methods leads to $\tilde{O}\left(\min\left(d^{9.5}\sqrt{T},\sqrt{d}T^{3/4}\right)\right)$ regret, that is significantly suboptimal in terms of d.
****
Asymptotics of Ridge Regression in Convolutional Models
Mojtaba Sahraee-Ardakan, Tung Mai, Anup Rao, Ryan A. Rossi, Sundeep Rangan, Alyson K Fletcher
Understanding generalization and estimation error of estimators for simple models such as linear and generalized linear models has attracted a lot of attention recently. This is in part due to an interesting observation made in machine learning community that highly over-parameterized neural networks achieve zero training error, and yet they are able to generalize well over the test samples. This phenomenon is captured by the so called double descent curve, where the generalization error starts decreasing again after the interpolation threshold. A series of recent works tried to explain such phenomenon for simple models. In this work, we analyze the asymptotics of estimation error in ridge estimators for convolutional linear models. These convolutional inverse problems, also known as deconvolution, naturally arise in different fields such as seismology, imaging, and acoustics among others. Our results hold for a large class of input distributions that include i.i.d. features as a special case. We derive exact formulae for estimation error of ridge estimators that hold in a certain high-dimensional regime. We show the double descent phenomenon in our experiments for convolutional models and show that our theoretical results match the experiments.
****
Momentum Residual Neural Networks
Michael E. Sander, Pierre Ablin, Mathieu Blondel, Gabriel Peyré
The training of deep residual neural networks (ResNets) with backpropagation has a memory cost that increases linearly with respect to the depth of the network. A simple way to circumvent this issue is to use reversible architectures. In this paper, we propose to change the forward rule of a ResNet by adding a momentum term. The resulting networks, momentum residual neural networks (MomentumNets), are invertible. Unlike previous invertible architectures, they can be used as a drop-in replacement for any existing ResNet block. We show that MomentumNets can be interpreted in the infinitesimal step size regime as second-order ordinary differential equations (ODEs) and exactly characterize how adding momentum progressively increases the representation capabilities of MomentumNets: they can learn any linear mapping up to a multiplicative factor, while ResNets cannot. In a learning to optimize setting, where convergence to a fixed point is required, we show theoretically and empirically that our method succeeds while existing invertible architectures fail. We show on CIFAR and ImageNet that MomentumNets have the same accuracy as ResNets, while having a much smaller memory footprint, and show that pre-trained MomentumNets are promising for fine-tuning models.
****
Meta-Learning Bidirectional Update Rules

Mark Sandler, Max Vladymyrov, Andrey Zhmoginov, Nolan Miller, Tom Madams, Andrew
 Jackson, Blaise Agüera Y Arcas
In this paper, we introduce a new type of generalized neural network where neuro
ns and synapses maintain multiple states. We show that classical gradient-based
backpropagation in neural networks can be seen as a special case of a two-state
network where one state is used for activations and another for gradients, with
update rules derived from the chain rule. In our generalized framework, networks
 have neither explicit notion of nor ever receive gradients. The synapses and ne
urons are updated using a bidirectional Hebb-style update rule parameterized by
a shared low-dimensional "genome". We show that such genomes can be meta-learned
 from scratch, using either conventional optimization techniques, or evolutionar
y strategies, such as CMA-ES. Resulting update rules generalize to unseen tasks
and train faster than gradient descent based optimizers for several standard com
puter vision and synthetic tasks.
****
Recomposing the Reinforcement Learning Building Blocks with Hypernetworks
Elad Sarafian, Shai Keynan, Sarit Kraus
The Reinforcement Learning (RL) building blocks, i.e. Q-functions and policy n
etworks, usually take elements from the cartesian product of two domains as inpu
t. In particular, the input of the Q-function is both the state and the action
, and in multi-task problems (Meta-RL) the policy can take a state and a context
. Standard architectures tend to ignore these variables’ underlying interpretati
ons and simply concatenate their features into a single vector. In this work, we
 argue that this choice may lead to poor gradient estimation in actor-critic alg
orithms and high variance learning steps in Meta-RL algorithms. To consider the
interaction between the input variables, we suggest using a Hypernetwork archite
cture where a primary network determines the weights of a conditional dynamic ne
twork. We show that this approach improves the gradient approximation and reduce
s the learning step variance, which both accelerates learning and improves the f
inal performance. We demonstrate a consistent improvement across different locom
otion tasks and different algorithms both in RL (TD3 and SAC) and in Meta-RL (MA
ML and PEARL).
****
Towards Understanding Learning in Neural Networks with Linear Teachers
Roei Sarussi, Alon Brutzkus, Amir Globerson
Can a neural network minimizing cross-entropy learn linearly separable data? Des
pite progress in the theory of deep learning, this question remains unsolved. He
re we prove that SGD globally optimizes this learning problem for a two-layer ne
twork with Leaky ReLU activations. The learned network can in principle be very
complex. However, empirical evidence suggests that it often turns out to be appr
oximately linear. We provide theoretical support for this phenomenon by proving
that if network weights converge to two weight clusters, this will imply an appr
oximately linear decision boundary. Finally, we show a condition on the optimiza
tion that leads to weight clustering. We provide empirical results that validate
 our theoretical analysis.
****
E(n) Equivariant Graph Neural Networks
V■■ctor Garcia Satorras, Emiel Hoogeboom, Max Welling
This paper introduces a new model to learn graph neural networks equivariant to
rotations, translations, reflections and permutations called E(n)-Equivariant Gr
aph Neural Networks (EGNNs). In contrast with existing methods, our work does no
t require computationally expensive higher-order representations in intermediate
 layers while it still achieves competitive or better performance. In addition,
whereas existing methods are limited to equivariance on 3 dimensional spaces, ou
r model is easily scaled to higher-dimensional spaces. We demonstrate the effect
iveness of our method on dynamical systems modelling, representation learning in
 graph autoencoders and predicting molecular properties.
****
A Representation Learning Perspective on the Importance of Train-Validation Spli
tting in Meta-Learning

Nikunj Saunshi, Arushi Gupta, Wei Hu
An effective approach in meta-learning is to utilize multiple “train tasks” to learn a good initialization for model parameters that can help solve unseen “test tasks” with very few samples by fine-tuning from this initialization. Although successful in practice, theoretical understanding of such methods is limited. This work studies an important aspect of these methods: splitting the data from each task into train (support) and validation (query) sets during meta-training. Inspired by recent work (Raghu et al., 2020), we view such meta-learning methods through the lens of representation learning and argue that the train-validation split encourages the learned representation to be {\em low-rank} without compromising on expressivity, as opposed to the non-splitting variant that encourages high-rank representations. Since sample efficiency benefits from low-rankness, the splitting strategy will require very few samples to solve unseen test tasks. We present theoretical results that formalize this idea for linear representation learning on a subspace meta-learning instance, and experimentally verify this practical benefit of splitting in simulations and on standard meta-learning benchmarks.
****
Low-Rank Sinkhorn Factorization
Meyer Scetbon, Marco Cuturi, Gabriel Peyré
Several recent applications of optimal transport (OT) theory to machine learning have relied on regularization, notably entropy and the Sinkhorn algorithm. Because matrix-vector products are pervasive in the Sinkhorn algorithm, several works have proposed to \textit{approximate} kernel matrices appearing in its iterations using low-rank factors. Another route lies instead in imposing low-nonnegative rank constraints on the feasible set of couplings considered in OT problems, with no approximations on cost nor kernel matrices. This route was first explored by \citet{forrow2018statistical}, who proposed an algorithm tailored for the squared Euclidean ground cost, using a proxy objective that can be solved through the machinery of regularized 2-Wasserstein barycenters. Building on this, we introduce in this work a generic approach that aims at solving, in full generality, the OT problem under low-nonnegative rank constraints with arbitrary costs. Our algorithm relies on an explicit factorization of low-rank couplings as a product of \textit{sub-coupling} factors linked by a common marginal; similar to an NMF approach, we alternatively updates these factors. We prove the non-asymptotic stationary convergence of this algorithm and illustrate its efficiency on benchmark experiments.
****
Linear Transformers Are Secretly Fast Weight Programmers
Imanol Schlag, Kazuki Irie, Jürgen Schmidhuber
We show the formal equivalence of linearised self-attention mechanisms and fast weight controllers from the early ’90s, where a slow neural net learns by gradient descent to program the fast weights of another net through sequences of elementary programming instructions which are additive outer products of self-invented activation patterns (today called keys and values). Such Fast Weight Programmers (FWPs) learn to manipulate the contents of a finite memory and dynamically interact with it. We infer a memory capacity limitation of recent linearised softmax attention variants, and replace the purely additive outer products by a delta rule-like programming instruction, such that the FWP can more easily learn to correct the current mapping from keys to values. The FWP also learns to compute dynamically changing learning rates. We also propose a new kernel function to linearise attention which balances simplicity and effectiveness. We conduct experiments on synthetic retrieval problems as well as standard machine translation and language modelling tasks which demonstrate the benefits of our methods.
****
Descending through a Crowded Valley - Benchmarking Deep Learning Optimizers
Robin M Schmidt, Frank Schneider, Philipp Hennig
Choosing the optimizer is considered to be among the most crucial design decisions in deep learning, and it is not an easy one. The growing literature now lists hundreds of optimization methods. In the absence of clear theoretical guidance

and conclusive empirical evidence, the decision is often made based on anecdotes. In this work, we aim to replace these anecdotes, if not with a conclusive ranking, then at least with evidence-backed heuristics. To do so, we perform an extensive, standardized benchmark of fifteen particularly popular deep learning optimizers while giving a concise overview of the wide range of possible choices. Analyzing more than 50,000 individual runs, we contribute the following three points: (i) Optimizer performance varies greatly across tasks. (ii) We observe that evaluating multiple optimizers with default parameters works approximately as well as tuning the hyperparameters of a single, fixed optimizer. (iii) While we cannot discern an optimization method clearly dominating across all tested tasks, we identify a significantly reduced subset of specific optimizers and parameter choices that generally lead to competitive results in our experiments: Adam remains a strong contender, with newer methods failing to significantly and consistently outperform it. Our open-sourced results are available as challenging and well-tuned baselines for more meaningful evaluations of novel optimization methods without requiring any further computational efforts.

****

Equivariant message passing for the prediction of tensorial properties and molecular spectra

Kristof Schütt, Oliver Unke, Michael Gastegger

Message passing neural networks have become a method of choice for learning on graphs, in particular the prediction of chemical properties and the acceleration of molecular dynamics studies. While they readily scale to large training data sets, previous approaches have proven to be less data efficient than kernel methods. We identify limitations of invariant representations as a major reason and extend the message passing formulation to rotationally equivariant representations. On this basis, we propose the polarizable atom interaction neural network (PaiNN) and improve on common molecule benchmarks over previous networks, while reducing model size and inference time. We leverage the equivariant atomwise representations obtained by PaiNN for the prediction of tensorial properties. Finally, we apply this to the simulation of molecular spectra, achieving speedups of 4-5 orders of magnitude compared to the electronic structure reference.

****

Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks

Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, Tom Goldstein

Data poisoning and backdoor attacks manipulate training data in order to cause models to fail during inference. A recent survey of industry practitioners found that data poisoning is the number one concern among threats ranging from model stealing to adversarial attacks. However, it remains unclear exactly how dangerous poisoning methods are and which ones are more effective considering that these methods, even ones with identical objectives, have not been tested in consistent or realistic settings. We observe that data poisoning and backdoor attacks are highly sensitive to variations in the testing setup. Moreover, we find that existing methods may not generalize to realistic settings. While these existing works serve as valuable prototypes for data poisoning, we apply rigorous tests to determine the extent to which we should fear them. In order to promote fair comparison in future work, we develop standardized benchmarks for data poisoning and backdoor attacks.

****

Connecting Sphere Manifolds Hierarchically for Regularization

Damien Scieur, Youngsung Kim

This paper considers classification problems with hierarchically organized classes. We force the classifier (hyperplane) of each class to belong to a sphere manifold, whose center is the classifier of its super-class. Then, individual sphere manifolds are connected based on their hierarchical relations. Our technique replaces the last layer of a neural network by combining a spherical fully-connected layer with a hierarchical layer. This regularization is shown to improve the performance of widely used deep neural network architectures (ResNet and DenseNet) on publicly available datasets (CIFAR100, CUB200, Stanford dogs, Stanford ca

rs, and Tiny-ImageNet).
****
Learning Intra-Batch Connections for Deep Metric Learning
Jenny Denise Seidenschwarz, Ismail Elezi, Laura Leal-Taixé
The goal of metric learning is to learn a function that maps samples to a lower-dimensional space where similar samples lie closer than dissimilar ones. Particularly, deep metric learning utilizes neural networks to learn such a mapping. Most approaches rely on losses that only take the relations between pairs or triplets of samples into account, which either belong to the same class or two different classes. However, these methods do not explore the embedding space in its entirety. To this end, we propose an approach based on message passing networks that takes all the relations in a mini-batch into account. We refine embedding vectors by exchanging messages among all samples in a given batch allowing the training process to be aware of its overall structure. Since not all samples are equally important to predict a decision boundary, we use an attention mechanism during message passing to allow samples to weigh the importance of each neighbor accordingly. We achieve state-of-the-art results on clustering and image retrieval on the CUB-200-2011, Cars196, Stanford Online Products, and In-Shop Clothes datasets. To facilitate further research, we make available the code and the models at https://github.com/dvl-tum/intra_batch_connections.
****
Top-k eXtreme Contextual Bandits with Arm Hierarchy
Rajat Sen, Alexander Rakhlin, Lexing Ying, Rahul Kidambi, Dean Foster, Daniel N Hill, Inderjit S. Dhillon
Motivated by modern applications, such as online advertisement and recommender systems, we study the top-k extreme contextual bandits problem, where the total number of arms can be enormous, and the learner is allowed to select k arms and observe all or some of the rewards for the chosen arms. We first propose an algorithm for the non-extreme realizable setting, utilizing the Inverse Gap Weighting strategy for selecting multiple arms. We show that our algorithm has a regret guarantee of $O(k\sqrt{(A-k+1)T \log (|F|T)})$, where A is the total number of arms and F is the class containing the regression function, while only requiring $\tilde{O}(A)$ computation per time step. In the extreme setting, where the total number of arms can be in the millions, we propose a practically-motivated arm hierarchy model that induces a certain structure in mean rewards to ensure statistical and computational efficiency. The hierarchical structure allows for an exponential reduction in the number of relevant arms for each context, thus resulting in a regret guarantee of $O(k\sqrt{(\log A-k+1)T \log (|F|T)})$. Finally, we implement our algorithm using a hierarchical linear function class and show superior performance with respect to well-known benchmarks on simulated bandit feedback experiments using extreme multi-label classification datasets. On a dataset with three million arms, our reduction scheme has an average inference time of only 7.9 milliseconds, which is a 100x improvement.
****
Pure Exploration and Regret Minimization in Matching Bandits
Flore Sentenac, Jialin Yi, Clement Calauzenes, Vianney Perchet, Milan Vojnovic
Finding an optimal matching in a weighted graph is a standard combinatorial problem. We consider its semi-bandit version where either a pair or a full matching is sampled sequentially. We prove that it is possible to leverage a rank-1 assumption on the adjacency matrix to reduce the sample complexity and the regret of off-the-shelf algorithms up to reaching a linear dependency in the number of vertices (up to to poly-log terms).
****
State Entropy Maximization with Random Encoders for Efficient Exploration
Younggyo Seo, Lili Chen, Jinwoo Shin, Honglak Lee, Pieter Abbeel, Kimin Lee
Recent exploration methods have proven to be a recipe for improving sample-efficiency in deep reinforcement learning (RL). However, efficient exploration in high-dimensional observation spaces still remains a challenge. This paper presents Random Encoders for Efficient Exploration (RE3), an exploration method that utilizes state entropy as an intrinsic reward. In order to estimate state entropy in

environments with high-dimensional observations, we utilize a k-nearest neighbor entropy estimator in the low-dimensional representation space of a convolutional encoder. In particular, we find that the state entropy can be estimated in a stable and compute-efficient manner by utilizing a randomly initialized encoder, which is fixed throughout training. Our experiments show that RE3 significantly improves the sample-efficiency of both model-free and model-based RL methods on locomotion and navigation tasks from DeepMind Control Suite and MiniGrid benchmarks. We also show that RE3 allows learning diverse behaviors without extrinsic rewards, effectively improving sample-efficiency in downstream tasks.
****
Online Submodular Resource Allocation with Applications to Rebalancing Shared Mobility Systems
Pier Giuseppe Sessa, Ilija Bogunovic, Andreas Krause, Maryam Kamgarpour
Motivated by applications in shared mobility, we address the problem of allocating a group of agents to a set of resources to maximize a cumulative welfare objective. We model the welfare obtainable from each resource as a monotone DR-submodular function which is a-priori unknown and can only be learned by observing the welfare of selected allocations. Moreover, these functions can depend on time-varying contextual information. We propose a distributed scheme to maximize the cumulative welfare by designing a repeated game among the agents, who learn to act via regret minimization. We propose two design choices for the game rewards based on upper confidence bounds built around the unknown welfare functions. We analyze them theoretically, bounding the gap between the cumulative welfare of the game and the highest cumulative welfare obtainable in hindsight. Finally, we evaluate our approach in a realistic case study of rebalancing a shared mobility system (i.e., positioning vehicles in strategic areas). From observed trip data, our algorithm gradually learns the users’ demand pattern and improves the overall system operation.
****
RRL: Resnet as representation for Reinforcement Learning
Rutav M Shah, Vikash Kumar
The ability to autonomously learn behaviors via direct interactions in uninstrumented environments can lead to generalist robots capable of enhancing productivity or providing care in unstructured settings like homes. Such uninstrumented settings warrant operations only using the robot’s proprioceptive sensor such as onboard cameras, joint encoders, etc which can be challenging for policy learning owing to the high dimensionality and partial observability issues. We propose RRL: Resnet as representation for Reinforcement Learning {–} a straightforward yet effective approach that can learn complex behaviors directly from proprioceptive inputs. RRL fuses features extracted from pre-trained Resnet into the standard reinforcement learning pipeline and delivers results comparable to learning directly from the state. In a simulated dexterous manipulation benchmark, where the state of the art methods fails to make significant progress, RRL delivers contact rich behaviors. The appeal of RRL lies in its simplicity in bringing together progress from the fields of Representation Learning, Imitation Learning, and Reinforcement Learning. Its effectiveness in learning behaviors directly from visual inputs with performance and sample efficiency matching learning directly from the state, even in complex high dimensional domains, is far from obvious.
****
Equivariant Networks for Pixelized Spheres
Mehran Shakerinava, Siamak Ravanbakhsh
Pixelizations of Platonic solids such as the cube and icosahedron have been widely used to represent spherical data, from climate records to Cosmic Microwave Background maps. Platonic solids have well-known global symmetries. Once we pixelize each face of the solid, each face also possesses its own local symmetries in the form of Euclidean isometries. One way to combine these symmetries is through a hierarchy. However, this approach does not adequately model the interplay between the two levels of symmetry transformations. We show how to model this interplay using ideas from group theory, identify the equivariant linear maps, and introduce equivariant padding that respects these symmetries. Deep networks that u

se these maps as their building blocks generalize gauge equivariant CNNs on pixe lized spheres. These deep networks achieve state-of-the-art results on semantic segmentation for climate data and omnidirectional image processing. Code is avai lable at https://git.io/JGiZA.
****

Personalized Federated Learning using Hypernetworks
Aviv Shamsian, Aviv Navon, Ethan Fetaya, Gal Chechik
Personalized federated learning is tasked with training machine learning models for multiple clients, each with its own data distribution. The goal is to train personalized models collaboratively while accounting for data disparities across clients and reducing communication costs. We propose a novel approach to this p roblem using hypernetworks, termed pFedHN for personalized Federated HyperNetwor ks. In this approach, a central hypernetwork model is trained to generate a set of models, one model for each client. This architecture provides effective param eter sharing across clients while maintaining the capacity to generate unique an d diverse personal models. Furthermore, since hypernetwork parameters are never transmitted, this approach decouples the communication cost from the trainable m odel size. We test pFedHN empirically in several personalized federated learning challenges and find that it outperforms previous methods. Finally, since hypern etworks share information across clients, we show that pFedHN can generalize bet ter to new clients whose distributions differ from any client observed during tr aining.
****

On the Power of Localized Perceptron for Label-Optimal Learning of Halfspaces wi th Adversarial Noise
Jie Shen
We study {\em online} active learning of homogeneous halfspaces in $\mathbb{R}^d$ with adversarial noise where the overall probability of a noisy label is const rained to be at most ν. Our main contribution is a Perceptron-like online ac tive learning algorithm that runs in polynomial time, and under the conditions t hat the marginal distribution is isotropic log-concave and $\nu = \Omega(\epsilon)$, where $\epsilon \in (0, 1)$ is the target error rate, our algorithm PAC lea rns the underlying halfspace with near-optimal label complexity of $\tilde{O}\big(d \cdot \polylog(\frac{1}{\epsilon})\big)$ and sample complexity of $\tilde{O}\big(\frac{d}{\epsilon} \big)$. Prior to this work, existing online algorithms d esigned for tolerating the adversarial noise are subject to either label complex ity polynomial in $\frac{1}{\epsilon}$, or suboptimal noise tolerance, or restri ctive marginal distributions. With the additional prior knowledge that the under lying halfspace is s-sparse, we obtain attribute-efficient label complexity of $\tilde{O}\big(s \cdot \polylog(d, \frac{1}{\epsilon}) \big)$ and sample compl exity of $\tilde{O}\big(\frac{s}{\epsilon} \cdot \polylog(d) \big)$. As an immed iate corollary, we show that under the agnostic model where no assumption is mad e on the noise rate ν, our active learner achieves an error rate of $O(OPT) + \epsilon$ with the same running time and label and sample complexity, where OPT is the best possible error rate achievable by any homogeneous halfspace.
****

Sample-Optimal PAC Learning of Halfspaces with Malicious Noise
Jie Shen
We study efficient PAC learning of homogeneous halfspaces in $\mathbb{R}^d$ in t he presence of malicious noise of Valiant (1985). This is a challenging noise mo del and only until recently has near-optimal noise tolerance bound been establis hed under the mild condition that the unlabeled data distribution is isotropic l og-concave. However, it remains unsettled how to obtain the optimal sample compl exity simultaneously. In this work, we present a new analysis for the algorithm of Awasthi et al. (2017) and show that it essentially achieves the near-optimal sample complexity bound of $\tilde{O}(d)$, improving the best known result of $\tilde{O}(d^2)$. Our main ingredient is a novel incorporation of a matrix Chernof f-type inequality to bound the spectrum of an empirical covariance matrix for we ll-behaved distributions, in conjunction with a careful exploration of the local ization schemes of Awasthi et al. (2017). We further extend the algorithm and an

alysis to the more general and stronger nasty noise model of Bshouty et al. (200
2), showing that it is still possible to achieve near-optimal noise tolerance an
d sample complexity in polynomial time.
****
Backdoor Scanning for Deep Neural Networks through K-Arm Optimization
Guangyu Shen, Yingqi Liu, Guanhong Tao, Shengwei An, Qiuling Xu, Siyuan Cheng, S
hiqing Ma, Xiangyu Zhang
Back-door attack poses a severe threat to deep learning systems. It injects hidd
en malicious behaviors to a model such that any input stamped with a special pat
tern can trigger such behaviors. Detecting back-door is hence of pressing need.
Many existing defense techniques use optimization to generate the smallest input
 pattern that forces the model to misclassify a set of benign inputs injected wi
th the pattern to a target label. However, the complexity is quadratic to the nu
mber of class labels such that they can hardly handle models with many classes.
Inspired by Multi-Arm Bandit in Reinforcement Learning, we propose a K-Arm optim
ization method for backdoor detection. By iteratively and stochastically selecti
ng the most promising labels for optimization with the guidance of an objective
function, we substantially reduce the complexity, allowing to handle models with
 many classes. Moreover, by iteratively refining the selection of labels to opti
mize, it substantially mitigates the uncertainty in choosing the right labels, i
mproving detection accuracy. At the time of submission, the evaluation of our me
thod on over 4000 models in the IARPA TrojAI competition from round 1 to the lat
est round 4 achieves top performance on the leaderboard. Our technique also supe
rsedes five state-of-the-art techniques in terms of accuracy and the scanning ti
me needed. The code of our work is available at https://github.com/PurduePAML/K-
ARM_Backdoor_Optimization
****
State Relevance for Off-Policy Evaluation
Simon P Shen, Yecheng Ma, Omer Gottesman, Finale Doshi-Velez
Importance sampling-based estimators for off-policy evaluation (OPE) are valued
for their simplicity, unbiasedness, and reliance on relatively few assumptions.
However, the variance of these estimators is often high, especially when traject
ories are of different lengths. In this work, we introduce Omitting-States-Irrel
evant-to-Return Importance Sampling (OSIRIS), an estimator which reduces varianc
e by strategically omitting likelihood ratios associated with certain states. We
 formalize the conditions under which OSIRIS is unbiased and has lower variance
than ordinary importance sampling, and we demonstrate these properties empirical
ly.
****
SparseBERT: Rethinking the Importance Analysis in Self-attention
Han Shi, Jiahui Gao, Xiaozhe Ren, Hang Xu, Xiaodan Liang, Zhenguo Li, James Tin-
Yau Kwok
Transformer-based models are popularly used in natural language processing (NLP)
. Its core component, self-attention, has aroused widespread interest. To unders
tand the self-attention mechanism, a direct method is to visualize the attention
 map of a pre-trained model. Based on the patterns observed, a series of efficie
nt Transformers with different sparse attention masks have been proposed. From a
 theoretical perspective, universal approximability of Transformer-based models
is also recently proved. However, the above understanding and analysis of self-a
ttention is based on a pre-trained model. To rethink the importance analysis in
self-attention, we study the significance of different positions in attention ma
trix during pre-training. A surprising result is that diagonal elements in the a
ttention map are the least important compared with other attention positions. We
 provide a proof showing that these diagonal elements can indeed be removed with
out deteriorating model performance. Furthermore, we propose a Differentiable At
tention Mask (DAM) algorithm, which further guides the design of the SparseBERT.
 Extensive experiments verify our interesting findings and illustrate the effect
 of the proposed algorithm.
****
Learning Gradient Fields for Molecular Conformation Generation

Chence Shi, Shitong Luo, Minkai Xu, Jian Tang
We study a fundamental problem in computational chemistry known as molecular conformation generation, trying to predict stable 3D structures from 2D molecular graphs. Existing machine learning approaches usually first predict distances between atoms and then generate a 3D structure satisfying the distances, where noise in predicted distances may induce extra errors during 3D coordinate generation. Inspired by the traditional force field methods for molecular dynamics simulation, in this paper, we propose a novel approach called ConfGF by directly estimating the gradient fields of the log density of atomic coordinates. The estimated gradient fields allow directly generating stable conformations via Langevin dynamics. However, the problem is very challenging as the gradient fields are roto-translation equivariant. We notice that estimating the gradient fields of atomic coordinates can be translated to estimating the gradient fields of interatomic distances, and hence develop a novel algorithm based on recent score-based generative models to effectively estimate these gradients. Experimental results across multiple tasks show that ConfGF outperforms previous state-of-the-art baselines by a significant margin.
****
Segmenting Hybrid Trajectories using Latent ODEs
Ruian Shi, Quaid Morris
Smooth dynamics interrupted by discontinuities are known as hybrid systems and arise commonly in nature. Latent ODEs allow for powerful representation of irregularly sampled time series but are not designed to capture trajectories arising from hybrid systems. Here, we propose the Latent Segmented ODE (LatSegODE), which uses Latent ODEs to perform reconstruction and changepoint detection within hybrid trajectories featuring jump discontinuities and switching dynamical modes. Where it is possible to train a Latent ODE on the smooth dynamical flows between discontinuities, we apply the pruned exact linear time (PELT) algorithm to detect changepoints where latent dynamics restart, thereby maximizing the joint probability of a piece-wise continuous latent dynamical representation. We propose usage of the marginal likelihood as a score function for PELT, circumventing the need for model-complexity-based penalization. The LatSegODE outperforms baselines in reconstructive and segmentation tasks including synthetic data sets of sine waves, Lotka Volterra dynamics, and UCI Character Trajectories.
****
Deeply-Debiased Off-Policy Interval Estimation
Chengchun Shi, Runzhe Wan, Victor Chernozhukov, Rui Song
Off-policy evaluation learns a target policy’s value with a historical dataset generated by a different behavior policy. In addition to a point estimate, many applications would benefit significantly from having a confidence interval (CI) that quantifies the uncertainty of the point estimate. In this paper, we propose a novel procedure to construct an efficient, robust, and flexible CI on a target policy’s value. Our method is justified by theoretical results and numerical experiments. A Python implementation of the proposed procedure is available at https://github.com/ RunzheStat/D2OPE.
****
GANMEX: One-vs-One Attributions using GAN-based Model Explainability
Sheng-Min Shih, Pin-Ju Tien, Zohar Karnin
Attribution methods have been shown as promising approaches for identifying key features that led to learned model predictions. While most existing attribution methods rely on a baseline input for performing feature perturbations, limited research has been conducted to address the baseline selection issues. Poor choices of baselines limit the ability of one-vs-one explanations for multi-class classifiers, which means the attribution methods were not able to explain why an input belongs to its original class but not the other specified target class. Achieving one-vs-one explanation is crucial when certain classes are more similar than others, e.g. two bird types among multiple animals, by focusing on key differentiating features rather than shared features across classes. In this paper, we present GANMEX, a novel approach applying Generative Adversarial Networks (GAN) by incorporating the to-be-explained classifier as part of the adversarial netwo

rks. Our approach effectively selects the baseline as the closest realistic sample belong to the target class, which allows attribution methods to provide true one-vs-one explanations. We showed that GANMEX baselines improved the saliency maps and led to stronger performance on multiple evaluation metrics over the existing baselines. Existing attribution results are known for being insensitive to model randomization, and we demonstrated that GANMEX baselines led to better outcome under the cascading randomization of the model.
****
Large-Scale Meta-Learning with Continual Trajectory Shifting
Jaewoong Shin, Hae Beom Lee, Boqing Gong, Sung Ju Hwang
Meta-learning of shared initialization parameters has shown to be highly effective in solving few-shot learning tasks. However, extending the framework to many-shot scenarios, which may further enhance its practicality, has been relatively overlooked due to the technical difficulties of meta-learning over long chains of inner-gradient steps. In this paper, we first show that allowing the meta-learners to take a larger number of inner gradient steps better captures the structure of heterogeneous and large-scale task distributions, thus results in obtaining better initialization points. Further, in order to increase the frequency of meta-updates even with the excessively long inner-optimization trajectories, we propose to estimate the required shift of the task-specific parameters with respect to the change of the initialization parameters. By doing so, we can arbitrarily increase the frequency of meta-updates and thus greatly improve the meta-level convergence as well as the quality of the learned initializations. We validate our method on a heterogeneous set of large-scale tasks, and show that the algorithm largely outperforms the previous first-order meta-learning methods in terms of both generalization performance and convergence, as well as multi-task learning and fine-tuning baselines.
****
AGENT: A Benchmark for Core Psychological Reasoning
Tianmin Shu, Abhishek Bhandwaldar, Chuang Gan, Kevin Smith, Shari Liu, Dan Gutfreund, Elizabeth Spelke, Joshua Tenenbaum, Tomer Ullman
For machine agents to successfully interact with humans in real-world settings, they will need to develop an understanding of human mental life. Intuitive psychology, the ability to reason about hidden mental variables that drive observable actions, comes naturally to people: even pre-verbal infants can tell agents from objects, expecting agents to act efficiently to achieve goals given constraints. Despite recent interest in machine agents that reason about other agents, it is not clear if such agents learn or hold the core psychology principles that drive human reasoning. Inspired by cognitive development studies on intuitive psychology, we present a benchmark consisting of a large dataset of procedurally generated 3D animations, AGENT (Action, Goal, Efficiency, coNstraint, uTility), structured around four scenarios (goal preferences, action efficiency, unobserved constraints, and cost-reward trade-offs) that probe key concepts of core intuitive psychology. We validate AGENT with human-ratings, propose an evaluation protocol emphasizing generalization, and compare two strong baselines built on Bayesian inverse planning and a Theory of Mind neural network. Our results suggest that to pass the designed tests of core intuitive psychology at human levels, a model must acquire or have built-in representations of how agents plan, combining utility computations and core knowledge of objects and physics.
****
Zoo-Tuning: Adaptive Transfer from A Zoo of Models
Yang Shu, Zhi Kou, Zhangjie Cao, Jianmin Wang, Mingsheng Long
With the development of deep networks on various large-scale datasets, a large zoo of pretrained models are available. When transferring from a model zoo, applying classic single-model-based transfer learning methods to each source model suffers from high computational cost and cannot fully utilize the rich knowledge in the zoo. We propose \emph{Zoo-Tuning} to address these challenges, which learns to adaptively transfer the parameters of pretrained models to the target task. With the learnable channel alignment layer and adaptive aggregation layer, Zoo-Tuning \emph{adaptively aggregates channel aligned pretrained parameters to deri

ve the target model}, which simultaneously promotes knowledge transfer and adapt
s source models to downstream tasks. The adaptive aggregation substantially redu
ces the computation cost at both training and inference. We further propose lite
 Zoo-Tuning with the temporal ensemble of batch average gating values to reduce
the storage cost at the inference time. We evaluate our approach on a variety of
 tasks, including reinforcement learning, image classification, and facial landm
ark detection. Experiment results demonstrate that the proposed adaptive transfe
r learning approach can more effectively and efficiently transfer knowledge from
 a zoo of models.
****
Aggregating From Multiple Target-Shifted Sources
Changjian Shui, Zijian Li, Jiaqi Li, Christian Gagné, Charles X Ling, Boyu Wang
Multi-source domain adaptation aims at leveraging the knowledge from multiple ta
sks for predicting a related target domain. Hence, a crucial aspect is to proper
ly combine different sources based on their relations. In this paper, we analyze
d the problem for aggregating source domains with different label distributions,
 where most recent source selection approaches fail. Our proposed algorithm diff
ers from previous approaches in two key ways: the model aggregates multiple sour
ces mainly through the similarity of semantic conditional distribution rather th
an marginal distribution; the model proposes a unified framework to select relev
ant sources for three popular scenarios, i.e., domain adaptation with limited la
bel on target domain, unsupervised domain adaptation and label partial unsupervi
sed domain adaption. We evaluate the proposed method through extensive experimen
ts. The empirical results significantly outperform the baselines.
****
Testing Group Fairness via Optimal Transport Projections
Nian Si, Karthyek Murthy, Jose Blanchet, Viet Anh Nguyen
We have developed a statistical testing framework to detect if a given machine l
earning classifier fails to satisfy a wide range of group fairness notions. Our
test is a flexible, interpretable, and statistically rigorous tool for auditing
whether exhibited biases are intrinsic to the algorithm or simply due to the ran
domness in the data. The statistical challenges, which may arise from multiple i
mpact criteria that define group fairness and which are discontinuous on model p
arameters, are conveniently tackled by projecting the empirical measure to the s
et of group-fair probability models using optimal transport. This statistic is e
fficiently computed using linear programming, and its asymptotic distribution is
 explicitly obtained. The proposed framework can also be used to test for compos
ite fairness hypotheses and fairness with multiple sensitive attributes. The opt
imal transport testing formulation improves interpretability by characterizing t
he minimal covariate perturbations that eliminate the bias observed in the audit
.
****
On Characterizing GAN Convergence Through Proximal Duality Gap
Sahil Sidheekh, Aroof Aimen, Narayanan C Krishnan
Despite the accomplishments of Generative Adversarial Networks (GANs) in modelin
g data distributions, training them remains a challenging task. A contributing f
actor to this difficulty is the non-intuitive nature of the GAN loss curves, whi
ch necessitates a subjective evaluation of the generated output to infer trainin
g progress. Recently, motivated by game theory, Duality Gap has been proposed as
 a domain agnostic measure to monitor GAN training. However, it is restricted to
 the setting when the GAN converges to a Nash equilibrium. But GANs need not alw
ays converge to a Nash equilibrium to model the data distribution. In this work,
 we extend the notion of duality gap to proximal duality gap that is applicable
to the general context of training GANs where Nash equilibria may not exist. We
show theoretically that the proximal duality gap can monitor the convergence of
GANs to a broader spectrum of equilibria that subsumes Nash equilibria. We also
theoretically establish the relationship between the proximal duality gap and th
e divergence between the real and generated data distributions for different GAN
 formulations. Our results provide new insights into the nature of GAN convergen
ce. Finally, we validate experimentally the usefulness of proximal duality gap f

or monitoring and influencing GAN training.
****
A Precise Performance Analysis of Support Vector Regression
Houssem Sifaou, Abla Kammoun, Mohamed-Slim Alouini
In this paper, we study the hard and soft support vector regression techniques applied to a set of n linear measurements of the form $y_i=\boldsymbol{\beta}_\star^{T}{\bf x}_i +n_i$ where $\boldsymbol{\beta}_\star$ is an unknown vector, $\left\{{\bf x}_i\right\}_{i=1}^n$ are the feature vectors and $\left\{{n}_i\right\}_{i=1}^n$ model the noise. Particularly, under some plausible assumptions on the statistical distribution of the data, we characterize the feasibility condition for the hard support vector regression in the regime of high dimensions and, when feasible, derive an asymptotic approximation for its risk. Similarly, we study the test risk for the soft support vector regression as a function of its parameters. Our results are then used to optimally tune the parameters intervening in the design of hard and soft support vector regression algorithms. Based on our analysis, we illustrate that adding more samples may be harmful to the test performance of support vector regression, while it is always beneficial when the parameters are optimally selected. Such a result reminds a similar phenomenon observed in modern learning architectures according to which optimally tuned architectures present a decreasing test performance curve with respect to the number of samples.
****
Directed Graph Embeddings in Pseudo-Riemannian Manifolds
Aaron Sim, Maciej L Wiatrak, Angus Brayne, Paidi Creed, Saee Paliwal
The inductive biases of graph representation learning algorithms are often encoded in the background geometry of their embedding space. In this paper, we show that general directed graphs can be effectively represented by an embedding model that combines three components: a pseudo-Riemannian metric structure, a non-trivial global topology, and a unique likelihood function that explicitly incorporates a preferred direction in embedding space. We demonstrate the representational capabilities of this method by applying it to the task of link prediction on a series of synthetic and real directed graphs from natural language applications and biology. In particular, we show that low-dimensional cylindrical Minkowski and anti-de Sitter spacetimes can produce equal or better graph representations than curved Riemannian manifolds of higher dimensions.
****
Collaborative Bayesian Optimization with Fair Regret
Rachael Hwee Ling Sim, Yehong Zhang, Bryan Kian Hsiang Low, Patrick Jaillet
Bayesian optimization (BO) is a popular tool for optimizing complex and costly-to-evaluate black-box objective functions. To further reduce the number of function evaluations, any party performing BO may be interested to collaborate with others to optimize the same objective function concurrently. To do this, existing BO algorithms have considered optimizing a batch of input queries in parallel and provided theoretical bounds on their cumulative regret reflecting inefficiency. However, when the objective function values are correlated with real-world rewards (e.g., money), parties may be hesitant to collaborate if they risk incurring larger cumulative regret (i.e., smaller real-world reward) than others. This paper shows that fairness and efficiency are both necessary for the collaborative BO setting. Inspired by social welfare concepts from economics, we propose a new notion of regret capturing these properties and a collaborative BO algorithm whose convergence rate can be theoretically guaranteed by bounding the new regret, both of which share an adjustable parameter for trading off between fairness vs. efficiency. We empirically demonstrate the benefits (e.g., increased fairness) of our algorithm using synthetic and real-world datasets.
****
Dynamic Planning and Learning under Recovering Rewards
David Simchi-Levi, Zeyu Zheng, Feng Zhu
Motivated by emerging applications such as live-streaming e-commerce, promotions and recommendations, we introduce a general class of multi-armed bandit problems that have the following two features: (i) the decision maker can pull and coll

ect rewards from at most K out of N different arms in each time period; (ii)
 the expected reward of an arm immediately drops after it is pulled, and then no
n-parametrically recovers as the idle time increases. With the objective of maxi
mizing expected cumulative rewards over T time periods, we propose, construct
and prove performance guarantees for a class of “Purely Periodic Policies”. For
the offline problem when all model parameters are known, our proposed policy obt
ains an approximation ratio that is at the order of $1-\mathcal O(1/\sqrt{K})$,
which is asymptotically optimal when K grows to infinity. For the online probl
em when the model parameters are unknown and need to be learned, we design an Up
per Confidence Bound (UCB) based policy that approximately has $\widetilde{\math
cal O}(N\sqrt{T})$ regret against the offline benchmark. Our framework and polic
y design may have the potential to be adapted into other offline planning and on
line learning applications with non-stationary and recovering rewards.
****
PopSkipJump: Decision-Based Attack for Probabilistic Classifiers
Carl-Johann Simon-Gabriel, Noman Ahmed Sheikh, Andreas Krause
Most current classifiers are vulnerable to adversarial examples, small input per
turbations that change the classification output. Many existing attack algorithm
s cover various settings, from white-box to black-box classifiers, but usually a
ssume that the answers are deterministic and often fail when they are not. We th
erefore propose a new adversarial decision-based attack specifically designed fo
r classifiers with probabilistic outputs. It is based on the HopSkipJump attack
by Chen et al. (2019), a strong and query efficient decision-based attack origin
ally designed for deterministic classifiers. Our P(robabilisticH)opSkipJump atta
ck adapts its amount of queries to maintain HopSkipJump’s original output qualit
y across various noise levels, while converging to its query efficiency as the n
oise level decreases. We test our attack on various noise models, including stat
e-of-the-art off-the-shelf randomized defenses, and show that they offer almost
no extra robustness to decision-based attacks. Code is available at https://gith
ub.com/cjsg/PopSkipJump.
****
Geometry of the Loss Landscape in Overparameterized Neural Networks: Symmetries
and Invariances
Berfin Simsek, François Ged, Arthur Jacot, Francesco Spadaro, Clement Hongler, W
ulfram Gerstner, Johanni Brea
We study how permutation symmetries in overparameterized multi-layer neural netw
orks generate ‘symmetry-induced’ critical points. Assuming a network with $ L $
layers of minimal widths $ r_1^*, \ldots, r_{L-1}^* $ reaches a zero-loss minimu
m at $ r_1^*! \cdots r_{L-1}^*! $ isolated points that are permutations of one a
nother, we show that adding one extra neuron to each layer is sufficient to conn
ect all these previously discrete minima into a single manifold. For a two-layer
 overparameterized network of width $ r^*+ h =: m $ we explicitly describe the m
anifold of global minima: it consists of $ T(r^*, m) $ affine subspaces of dimen
sion at least $ h $ that are connected to one another. For a network of width $m
$, we identify the number $G(r,m)$ of affine subspaces containing only symmetry-
induced critical points that are related to the critical points of a smaller net
work of width $r
Cite this Paper

 BibTeX

```
@InProceedings{pmlr-v139-simsek21a,
  title = ■ {Geometry of the Loss Landscape in Overparameterized Neural Networks
: Symmetries and Invariances},
  author =       {Simsek, Berfin and Ged, Fran{\c{c}}ois and Jacot, Arthur and S
padaro, Francesco and Hongler, Clement and Gerstner, Wulfram and Brea, Johanni},
  booktitle = ■ {Proceedings of the 38th International Conference on Machine Lea
rning},
  pages = ■ {9722--9732},
  year = ■ {2021},
```

```
  editor = ■ {Meila, Marina and Zhang, Tong},
  volume = ■ {139},
  series = ■ {Proceedings of Machine Learning Research},
  month = ■ {18--24 Jul},
  publisher =    {PMLR},
  pdf = ■ {http://proceedings.mlr.press/v139/simsek21a/simsek21a.pdf},
  url = ■ {https://proceedings.mlr.press/v139/simsek21a.html},
  abstract = ■ {We study how permutation symmetries in overparameterized multi-l
ayer neural networks generate ‘symmetry-induced’ critical points. Assuming a net
work with $ L $ layers of minimal widths $ r_1^*, \ldots, r_{L-1}^* $ reaches a 
zero-loss minimum at $ r_1^*! \cdots r_{L-1}^*! $ isolated points that are permu
tations of one another, we show that adding one extra neuron to each layer is su
fficient to connect all these previously discrete minima into a single manifold.
 For a two-layer overparameterized network of width $ r^*+ h =: m $ we explicitl
y describe the manifold of global minima: it consists of $ T(r^*, m) $ affine su
bspaces of dimension at least $ h $ that are connected to one another. For a net
work of width $m$, we identify the number $G(r,m)$ of affine subspaces containin
g only symmetry-induced critical points that are related to the critical points 
of a smaller network of width $r
```
Copy to ClipboardDownload
 Endnote
```
%0 Conference Paper
%T Geometry of the Loss Landscape in Overparameterized Neural Networks: Symmetri
es and Invariances
%A Berfin Simsek
%A François Ged
%A Arthur Jacot
%A Francesco Spadaro
%A Clement Hongler
%A Wulfram Gerstner
%A Johanni Brea
%B Proceedings of the 38th International Conference on Machine Learning
%C Proceedings of Machine Learning Research
%D 2021
%E Marina Meila
%E Tong Zhang■
%F pmlr-v139-simsek21a
%I PMLR
%P 9722--9732
%U https://proceedings.mlr.press/v139/simsek21a.html
%V 139
%X We study how permutation symmetries in overparameterized multi-layer neural n
etworks generate ‘symmetry-induced’ critical points. Assuming a network with $ L
 $ layers of minimal widths $ r_1^*, \ldots, r_{L-1}^* $ reaches a zero-loss min
imum at $ r_1^*! \cdots r_{L-1}^*! $ isolated points that are permutations of on
e another, we show that adding one extra neuron to each layer is sufficient to c
onnect all these previously discrete minima into a single manifold. For a two-la
yer overparameterized network of width $ r^*+ h =: m $ we explicitly describe th
e manifold of global minima: it consists of $ T(r^*, m) $ affine subspaces of di
mension at least $ h $ that are connected to one another. For a network of width
 $m$, we identify the number $G(r,m)$ of affine subspaces containing only symmet
ry-induced critical points that are related to the critical points of a smaller 
network of width $r
```
Copy to ClipboardDownload
 APA

Simsek, B., Ged, F., Jacot, A., Spadaro, F., Hongler, C., Gerstner, W. & Brea, J.. (2021). Geometry of the Loss Landscape in Overparameterized Neural Networks: Symmetries and Invariances. Proceedings of the 38th International Conference on Machine Learning, in Proceedings of Machine Learning Research 139:9722-9732 Avai lable from https://proceedings.mlr.press/v139/simsek21a.html.

Copy to ClipboardDownload

Related Material

Download PDF
Supplementary PDF
****
Flow-based Attribution in Graphical Models: A Recursive Shapley Approach
Raghav Singal, George Michailidis, Hoiyi Ng
We study the attribution problem in a graphical model, wherein the objective is to quantify how the effect of changes at the source nodes propagates through the graph. We develop a model-agnostic flow-based attribution method, called recursive Shapley value (RSV). RSV generalizes a number of existing node-based methods and uniquely satisfies a set of flow-based axioms. In addition to admitting a natural characterization for linear models and facilitating mediation analysis for non-linear models, RSV satisfies a mix of desirable properties discussed in the recent literature, including implementation invariance, sensitivity, monotonicity, and affine scale invariance.
****
Structured World Belief for Reinforcement Learning in POMDP
Gautam Singh, Skand Peri, Junghyun Kim, Hyunseok Kim, Sungjin Ahn
Object-centric world models provide structured representation of the scene and can be an important backbone in reinforcement learning and planning. However, existing approaches suffer in partially-observable environments due to the lack of belief states. In this paper, we propose Structured World Belief, a model for learning and inference of object-centric belief states. Inferred by Sequential Monte Carlo (SMC), our belief states provide multiple object-centric scene hypotheses. To synergize the benefits of SMC particles with object representations, we also propose a new object-centric dynamics model that considers the inductive bias of object permanence. This enables tracking of object states even when they are invisible for a long time. To further facilitate object tracking in this regime, we allow our model to attend flexibly to any spatial location in the image which was restricted in previous models. In experiments, we show that object-centric belief provides a more accurate and robust performance for filtering and generation. Furthermore, we show the efficacy of structured world belief in improving the performance of reinforcement learning, planning and supervised reasoning.
****
Skew Orthogonal Convolutions
Sahil Singla, Soheil Feizi
Training convolutional neural networks with a Lipschitz constraint under the l_{2} norm is useful for provable adversarial robustness, interpretable gradients, stable training, etc. While 1-Lipschitz networks can be designed by imposing a 1-Lipschitz constraint on each layer, training such networks requires each layer to be gradient norm preserving (GNP) to prevent gradients from vanishing. However, existing GNP convolutions suffer from slow training, lead to significant reduction in accuracy and provide no guarantees on their approximations. In this work, we propose a GNP convolution layer called \textbf{S}kew \textbf{O}rthogonal \textbf{C}onvolution (SOC) that uses the following mathematical property: when a matrix is {\it Skew-Symmetric}, its exponential function is an {\it orthogonal} matrix. To use this property, we first construct a convolution filter whose Jacobian is Skew-Symmetric. Then, we use the Taylor series expansion of the Jacobian exponential to construct the SOC layer that is orthogonal. To efficiently implement SOC, we keep a finite number of terms from the Taylor series and provide a provable guarantee on the approximation error. Our experiments on CIFAR-10 and CIFAR-100 show that SOC allows us to train provably Lipschitz, large convolutional neural networks significantly faster than prior works while achieving signif

icant improvements for both standard and certified robust accuracies.
****
Multi-Task Reinforcement Learning with Context-based Representations
Shagun Sodhani, Amy Zhang, Joelle Pineau
https://drive.google.com/file/d/1lRV72XaKoxZjgQrLXBJhsM82x54_1Vc4/view?usp=sharing
****
Shortest-Path Constrained Reinforcement Learning for Sparse Reward Tasks
Sungryull Sohn, Sungtae Lee, Jongwook Choi, Harm H Van Seijen, Mehdi Fatemi, Honglak Lee
We propose the k-Shortest-Path (k-SP) constraint: a novel constraint on the agent's trajectory that improves the sample efficiency in sparse-reward MDPs. We show that any optimal policy necessarily satisfies the k-SP constraint. Notably, the k-SP constraint prevents the policy from exploring state-action pairs along the non-k-SP trajectories (e.g., going back and forth). However, in practice, excluding state-action pairs may hinder the convergence of RL algorithms. To overcome this, we propose a novel cost function that penalizes the policy violating SP constraint, instead of completely excluding it. Our numerical experiment in a tabular RL setting demonstrates that the SP-constraint can significantly reduce the trajectory space of policy. As a result, our constraint enables more sample efficient learning by suppressing redundant exploration and exploitation. Our experiments on MiniGrid, DeepMind Lab, Atari, and Fetch show that the proposed method significantly improves proximal policy optimization (PPO) and outperforms existing novelty-seeking exploration methods including count-based exploration even in continuous control tasks, indicating that it improves the sample efficiency by preventing the agent from taking redundant actions.
****
Accelerating Feedforward Computation via Parallel Nonlinear Equation Solving
Yang Song, Chenlin Meng, Renjie Liao, Stefano Ermon
Feedforward computation, such as evaluating a neural network or sampling from an autoregressive model, is ubiquitous in machine learning. The sequential nature of feedforward computation, however, requires a strict order of execution and cannot be easily accelerated with parallel computing. To enable parallelization, we frame the task of feedforward computation as solving a system of nonlinear equations. We then propose to find the solution using a Jacobi or Gauss-Seidel fixed-point iteration method, as well as hybrid methods of both. Crucially, Jacobi updates operate independently on each equation and can be executed in parallel. Our method is guaranteed to give exactly the same values as the original feedforward computation with a reduced (or equal) number of parallelizable iterations, and hence reduced time given sufficient parallel computing power. Experimentally, we demonstrate the effectiveness of our approach in accelerating (i) backpropagation of RNNs, (ii) evaluation of DenseNets, and (iii) autoregressive sampling of MADE and PixelCNN++, with speedup factors between 2.1 and 26 under various settings.
****
PC-MLP: Model-based Reinforcement Learning with Policy Cover Guided Exploration
Yuda Song, Wen Sun
Model-based Reinforcement Learning (RL) is a popular learning paradigm due to its potential sample efficiency compared to model-free RL. However, existing empirical model-based RL approaches lack the ability to explore. This work studies a computationally and statistically efficient model-based algorithm for both Kernelized Nonlinear Regulators (KNR) and linear Markov Decision Processes (MDPs). For both models, our algorithm guarantees polynomial sample complexity and only uses access to a planning oracle. Experimentally, we first demonstrate the flexibility and the efficacy of our algorithm on a set of exploration challenging control tasks where existing empirical model-based RL approaches completely fail. We then show that our approach retains excellent performance even in common dense reward control benchmarks that do not require heavy exploration.
****
Fast Sketching of Polynomial Kernels of Polynomial Degree

Zhao Song, David Woodruff, Zheng Yu, Lichen Zhang
Kernel methods are fundamental in machine learning, and faster algorithms for ke
rnel approximation provide direct speedups for many core tasks in machine learni
ng. The polynomial kernel is especially important as other kernels can often be
approximated by the polynomial kernel via a Taylor series expansion. Recent tech
niques in oblivious sketching reduce the dependence in the running time on the d
egree q of the polynomial kernel from exponential to polynomial, which is usef
ul for the Gaussian kernel, for which q can be chosen to be polylogarithmic. H
owever, for more slowly growing kernels, such as the neural tangent and arc cosi
ne kernels, q needs to be polynomial, and previous work incurs a polynomial fa
ctor slowdown in the running time. We give a new oblivious sketch which greatly
improves upon this running time, by removing the dependence on q in the leadin
g order term. Combined with a novel sampling scheme, we give the fastest algorit
hms for approximating a large family of slow-growing kernels.
****
Variance Reduction via Primal-Dual Accelerated Dual Averaging for Nonsmooth Conv
ex Finite-Sums
Chaobing Song, Stephen J Wright, Jelena Diakonikolas
Structured nonsmooth convex finite-sum optimization appears in many machine lear
ning applications, including support vector machines and least absolute deviatio
n. For the primal-dual formulation of this problem, we propose a novel algorithm
called \emph{Variance Reduction via Primal-Dual Accelerated Dual Averaging (\vr
pda)}. In the nonsmooth and general convex setting, \vrpda has the overall compl
exity $O(nd\log\min \{1/\epsilon, n\} + d/\epsilon)$ in terms of the primal-dua
l gap, where n denotes the number of samples, d the dimension of the primal
variables, and ϵ the desired accuracy. In the nonsmooth and strongly co
nvex setting, the overall complexity of \vrpda becomes $O(nd\log\min\{1/\epsilon
, n\} + d/\sqrt{\epsilon})$ in terms of both the primal-dual gap and the distanc
e between iterate and optimal solution. Both these results for \vrpda improve si
gnificantly on state-of-the-art complexity estimates—which are $O(nd\log \min\{1
/\epsilon, n\} + \sqrt{n}d/\epsilon)$ for the nonsmooth and general convex setti
ng and $O(nd\log \min\{1/\epsilon, n\} + \sqrt{n}d/\sqrt{\epsilon})$ for the non
smooth and strongly convex setting—with a simpler and more straightforward algor
ithm and analysis. Moreover, both complexities are better than \emph{lower} boun
ds for general convex finite-sum optimization, because our approach makes use of
additional, commonly occurring structure. Numerical experiments reveal competit
ive performance of \vrpda compared to state-of-the-art approaches.
****
Oblivious Sketching-based Central Path Method for Linear Programming
Zhao Song, Zheng Yu
In this work, we propose a sketching-based central path method for solving linea
r programmings, whose running time matches the state of the art results [Cohen,
Lee, Song STOC 19; Lee, Song, Zhang COLT 19]. Our method opens up the iterations
of the central path method and deploys an "iterate and sketch" approach towards
the problem by introducing a new coordinate-wise embedding technique, which may
be of independent interest. Compare to previous methods, the work [Cohen, Lee,
Song STOC 19] enjoys feasibility while being non-oblivious, and [Lee, Song, Zhan
g COLT 19] is oblivious but infeasible, and relies on dense sketching
matrices such as subsampled randomized Hadamard/Fourier transform matrices. Our
method enjoys the benefits of being both oblivious and feasible, and can use $\
mathit{sparse}$ sketching matrix [Nelson, Nguyen FOCS 13] to speed up the online
matrix-vector multiplication. Our framework for solving LP naturally generalize
s to a broader class of convex optimization problems including empirical risk mi
nimization.
****
Causal Curiosity: RL Agents Discovering Self-supervised Experiments for Causal R
epresentation Learning
Sumedh A Sontakke, Arash Mehrjou, Laurent Itti, Bernhard Schölkopf
Humans show an innate ability to learn the regularities of the world through int
eraction. By performing experiments in our environment, we are able to discern t

he causal factors of variation and infer how they affect the dynamics of our world. Analogously, here we attempt to equip reinforcement learning agents with the ability to perform experiments that facilitate a categorization of the rolled-out trajectories, and to subsequently infer the causal factors of the environment in a hierarchical manner. We introduce a novel intrinsic reward, called causal curiosity, and show that it allows our agents to learn optimal sequences of actions, and to discover causal factors in the dynamics. The learned behavior allows the agent to infer a binary quantized representation for the ground-truth causal factors in every environment. Additionally, we find that these experimental behaviors are semantically meaningful (e.g., to differentiate between heavy and light blocks, our agents learn to lift them), and are learnt in a self-supervised manner with approximately 2.5 times less data than conventional supervised planners. We show that these behaviors can be re-purposed and fine-tuned (e.g., from lifting to pushing or other downstream tasks). Finally, we show that the knowledge of causal factor representations aids zero-shot learning for more complex tasks.
****
Decomposed Mutual Information Estimation for Contrastive Representation Learning
Alessandro Sordoni, Nouha Dziri, Hannes Schulz, Geoff Gordon, Philip Bachman, Remi Tachet Des Combes
Recent contrastive representation learning methods rely on estimating mutual information (MI) between multiple views of an underlying context. E.g., we can derive multiple views of a given image by applying data augmentation, or we can split a sequence into views comprising the past and future of some step in the sequence. Contrastive lower bounds on MI are easy to optimize, but have a strong underestimation bias when estimating large amounts of MI. We propose decomposing the full MI estimation problem into a sum of smaller estimation problems by splitting one of the views into progressively more informed subviews and by applying the chain rule on MI between the decomposed views. This expression contains a sum of unconditional and conditional MI terms, each measuring modest chunks of the total MI, which facilitates approximation via contrastive bounds. To maximize the sum, we formulate a contrastive lower bound on the conditional MI which can be approximated efficiently. We refer to our general approach as Decomposed Estimation of Mutual Information (DEMI). We show that DEMI can capture a larger amount of MI than standard non-decomposed contrastive bounds in a synthetic setting, and learns better representations in a vision domain and for dialogue generation.
****
Decoupling Representation Learning from Reinforcement Learning
Adam Stooke, Kimin Lee, Pieter Abbeel, Michael Laskin
In an effort to overcome limitations of reward-driven feature learning in deep reinforcement learning (RL) from images, we propose decoupling representation learning from policy learning. To this end, we introduce a new unsupervised learning (UL) task, called Augmented Temporal Contrast (ATC), which trains a convolutional encoder to associate pairs of observations separated by a short time difference, under image augmentations and using a contrastive loss. In online RL experiments, we show that training the encoder exclusively using ATC matches or outperforms end-to-end RL in most environments. Additionally, we benchmark several leading UL algorithms by pre-training encoders on expert demonstrations and using them, with weights frozen, in RL agents; we find that agents using ATC-trained encoders outperform all others. We also train multi-task encoders on data from multiple environments and show generalization to different downstream RL tasks. Finally, we ablate components of ATC, and introduce a new data augmentation to enable replay of (compressed) latent images from pre-trained encoders when RL requires augmentation. Our experiments span visually diverse RL benchmarks in DeepMind Control, DeepMind Lab, and Atari, and our complete code is available at \url{https://github.com/astooke/rlpyt/tree/master/rlpyt/ul}.
****
K-shot NAS: Learnable Weight-Sharing for NAS with K-shot Supernets
Xiu Su, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, Changshui Zhang, Chang Xu
In one-shot weight sharing for NAS, the weights of each operation (at each layer

) are supposed to be identical for all architectures (paths) in the supernet. Ho
wever, this rules out the possibility of adjusting operation weights to cater fo
r different paths, which limits the reliability of the evaluation results. In th
is paper, instead of counting on a single supernet, we introduce K-shot supern
ets and take their weights for each operation as a dictionary. The operation wei
ght for each path is represented as a convex combination of items in a dictionar
y with a simplex code. This enables a matrix approximation of the stand-alone we
ight matrix with a higher rank ($K>1$). A \textit{simplex-net} is introduced to
produce architecture-customized code for each path. As a result, all paths can a
daptively learn how to share weights in the K-shot supernets and acquire corre
sponding weights for better evaluation. K-shot supernets and simplex-net can b
e iteratively trained, and we further extend the search to the channel dimension
. Extensive experiments on benchmark datasets validate that K-shot NAS significa
ntly improves the evaluation accuracy of paths and thus brings in impressive per
formance improvements.
****
More Powerful and General Selective Inference for Stepwise Feature Selection usi
ng Homotopy Method
Kazuya Sugiyama, Vo Nguyen Le Duy, Ichiro Takeuchi
Conditional selective inference (SI) has been actively studied as a new statisti
cal inference framework for data-driven hypotheses. The basic idea of conditiona
l SI is to make inferences conditional on the selection event characterized by a
 set of linear and/or quadratic inequalities. Conditional SI has been mainly stu
died in the context of feature selection such as stepwise feature selection (SFS
). The main limitation of the existing conditional SI methods is the loss of pow
er due to over-conditioning, which is required for computational tractability. I
n this study, we develop a more powerful and general conditional SI method for S
FS using the homotopy method which enables us to overcome this limitation. The h
omotopy-based SI is especially effective for more complicated feature selection
algorithms. As an example, we develop a conditional SI method for forward-backwa
rd SFS with AIC-based stopping criteria and show that it is not adversely affect
ed by the increased complexity of the algorithm. We conduct several experiments
to demonstrate the effectiveness and efficiency of the proposed method.
****
Not All Memories are Created Equal: Learning to Forget by Expiring
Sainbayar Sukhbaatar, Da Ju, Spencer Poff, Stephen Roller, Arthur Szlam, Jason W
eston, Angela Fan
Attention mechanisms have shown promising results in sequence modeling tasks tha
t require long-term memory. Recent work investigated mechanisms to reduce the co
mputational cost of preserving and storing memories. However, not all content in
 the past is equally important to remember. We propose Expire-Span, a method tha
t learns to retain the most important information and expire the irrelevant info
rmation. This forgetting of memories enables Transformers to scale to attend ove
r tens of thousands of previous timesteps efficiently, as not all states from pr
evious timesteps are preserved. We demonstrate that Expire-Span can help models
identify and retain critical information and show it can achieve strong performa
nce on reinforcement learning tasks specifically designed to challenge this func
tionality. Next, we show that Expire-Span can scale to memories that are tens of
 thousands in size, setting a new state of the art on incredibly long context ta
sks such as character-level language modeling and a frame-by-frame moving object
s task. Finally, we analyze the efficiency of Expire-Span compared to existing a
pproaches and demonstrate that it trains faster and uses less memory.
****
Nondeterminism and Instability in Neural Network Optimization
Cecilia Summers, Michael J. Dinneen
Nondeterminism in neural network optimization produces uncertainty in performanc
e, making small improvements difficult to discern from run-to-run variability. W
hile uncertainty can be reduced by training multiple model copies, doing so is t
ime-consuming, costly, and harms reproducibility. In this work, we establish an
experimental protocol for understanding the effect of optimization nondeterminis

m on model diversity, allowing us to isolate the effects of a variety of sources of nondeterminism. Surprisingly, we find that all sources of nondeterminism have similar effects on measures of model diversity. To explain this intriguing fact, we identify the instability of model training, taken as an end-to-end procedure, as the key determinant. We show that even one-bit changes in initial parameters result in models converging to vastly different values. Last, we propose two approaches for reducing the effects of instability on run-to-run variability.
****
AutoSampling: Search for Effective Data Sampling Schedules
Ming Sun, Haoxuan Dou, Baopu Li, Junjie Yan, Wanli Ouyang, Lei Cui
Data sampling acts as a pivotal role in training deep learning models. However, an effective sampling schedule is difficult to learn due to its inherent high-dimension as a hyper-parameter. In this paper, we propose an AutoSampling method to automatically learn sampling schedules for model training, which consists of the multi-exploitation step aiming for optimal local sampling schedules and the exploration step for the ideal sampling distribution. More specifically, we achieve sampling schedule search with shortened exploitation cycle to provide enough supervision. In addition, we periodically estimate the sampling distribution from the learned sampling schedules and perturb it to search in the distribution space. The combination of two searches allows us to learn a robust sampling schedule. We apply our AutoSampling method to a variety of image classification tasks illustrating the effectiveness of the proposed method.
****
What Makes for End-to-End Object Detection?
Peize Sun, Yi Jiang, Enze Xie, Wenqi Shao, Zehuan Yuan, Changhu Wang, Ping Luo
Object detection has recently achieved a breakthrough for removing the last one non-differentiable component in the pipeline, Non-Maximum Suppression (NMS), and building up an end-to-end system. However, what makes for its one-to-one prediction has not been well understood. In this paper, we first point out that one-to-one positive sample assignment is the key factor, while, one-to-many assignment in previous detectors causes redundant predictions in inference. Second, we surprisingly find that even training with one-to-one assignment, previous detectors still produce redundant predictions. We identify that classification cost in matching cost is the main ingredient: (1) previous detectors only consider location cost, (2) by additionally introducing classification cost, previous detectors immediately produce one-to-one prediction during inference. We introduce the concept of score gap to explore the effect of matching cost. Classification cost enlarges the score gap by choosing positive samples as those of highest score in the training iteration and reducing noisy positive samples brought by only location cost. Finally, we demonstrate the advantages of end-to-end object detection on crowded scenes.
****
DFAC Framework: Factorizing the Value Function via Quantile Mixture for Multi-Agent Distributional Q-Learning
Wei-Fang Sun, Cheng-Kuang Lee, Chun-Yi Lee
In fully cooperative multi-agent reinforcement learning (MARL) settings, the environments are highly stochastic due to the partial observability of each agent and the continuously changing policies of the other agents. To address the above issues, we integrate distributional RL and value function factorization methods by proposing a Distributional Value Function Factorization (DFAC) framework to generalize expected value function factorization methods to their distributional variants. DFAC extends the individual utility functions from deterministic variables to random variables, and models the quantile function of the total return as a quantile mixture. To validate DFAC, we demonstrate DFAC’s ability to factorize a simple two-step matrix game with stochastic rewards and perform experiments on all Super Hard tasks of StarCraft Multi-Agent Challenge, showing that DFAC is able to outperform expected value function factorization baselines.
****
Scalable Variational Gaussian Processes via Harmonic Kernel Decomposition
Shengyang Sun, Jiaxin Shi, Andrew Gordon Gordon Wilson, Roger B Grosse

We introduce a new scalable variational Gaussian process approximation which provides a high fidelity approximation while retaining general applicability. We propose the harmonic kernel decomposition (HKD), which uses Fourier series to decompose a kernel as a sum of orthogonal kernels. Our variational approximation exploits this orthogonality to enable a large number of inducing points at a low computational cost. We demonstrate that, on a range of regression and classification problems, our approach can exploit input space symmetries such as translations and reflections, and it significantly outperforms standard variational methods in scalability and accuracy. Notably, our approach achieves state-of-the-art results on CIFAR-10 among pure GP models.
****
Reasoning Over Virtual Knowledge Bases With Open Predicate Relations
Haitian Sun, Patrick Verga, Bhuwan Dhingra, Ruslan Salakhutdinov, William W Cohen
We present the Open Predicate Query Language (OPQL); a method for constructing a virtual KB (VKB) trained entirely from text. Large Knowledge Bases (KBs) are indispensable for a wide-range of industry applications such as question answering and recommendation. Typically, KBs encode world knowledge in a structured, readily accessible form derived from laborious human annotation efforts. Unfortunately, while they are extremely high precision, KBs are inevitably highly incomplete and automated methods for enriching them are far too inaccurate. Instead, OPQL constructs a VKB by encoding and indexing a set of relation mentions in a way that naturally enables reasoning and can be trained without any structured supervision. We demonstrate that OPQL outperforms prior VKB methods on two different KB reasoning tasks and, additionally, can be used as an external memory integrated into a language model (OPQL-LM) leading to improvements on two open-domain question answering tasks.
****
PAC-Learning for Strategic Classification
Ravi Sundaram, Anil Vullikanti, Haifeng Xu, Fan Yao
The study of strategic or adversarial manipulation of testing data to fool a classifier has attracted much recent attention. Most previous works have focused on two extreme situations where any testing data point either is completely adversarial or always equally prefers the positive label. In this paper, we generalize both of these through a unified framework for strategic classification and introduce the notion of strategic VC-dimension (SVC) to capture the PAC-learnability in our general strategic setup. SVC provably generalizes the recent concept of adversarial VC-dimension (AVC) introduced by Cullina et al. (2018). We instantiate our framework for the fundamental strategic linear classification problem. We fully characterize: (1) the statistical learnability of linear classifiers by pinning down its SVC; (2) it’s computational tractability by pinning down the complexity of the empirical risk minimization problem. Interestingly, the SVC of linear classifiers is always upper bounded by its standard VC-dimension. This characterization also strictly generalizes the AVC bound for linear classifiers in (Cullina et al., 2018).
****
Reinforcement Learning for Cost-Aware Markov Decision Processes
Wesley Suttle, Kaiqing Zhang, Zhuoran Yang, Ji Liu, David Kraemer
Ratio maximization has applications in areas as diverse as finance, reward shaping for reinforcement learning (RL), and the development of safe artificial intelligence, yet there has been very little exploration of RL algorithms for ratio maximization. This paper addresses this deficiency by introducing two new, model-free RL algorithms for solving cost-aware Markov decision processes, where the goal is to maximize the ratio of long-run average reward to long-run average cost. The first algorithm is a two-timescale scheme based on relative value iteration (RVI) Q-learning and the second is an actor-critic scheme. The paper proves almost sure convergence of the former to the globally optimal solution in the tabular case and almost sure convergence of the latter under linear function approximation for the critic. Unlike previous methods, the two algorithms provably converge for general reward and cost functions under suitable conditions. The paper

also provides empirical results demonstrating promising performance and lending strong support to the theoretical results.
****

Model-Targeted Poisoning Attacks with Provable Convergence

Fnu Suya, Saeed Mahloujifar, Anshuman Suri, David Evans, Yuan Tian

In a poisoning attack, an adversary who controls a small fraction of the training data attempts to select that data, so a model is induced that misbehaves in a particular way. We consider poisoning attacks against convex machine learning models and propose an efficient poisoning attack designed to induce a model specified by the adversary. Unlike previous model-targeted poisoning attacks, our attack comes with provable convergence to any attainable target model. We also provide a lower bound on the minimum number of poisoning points needed to achieve a given target model. Our method uses online convex optimization and finds poisoning points incrementally. This provides more flexibility than previous attacks which require an a priori assumption about the number of poisoning points. Our attack is the first model-targeted poisoning attack that provides provable convergence for convex models. In our experiments, it either exceeds or matches state-of-the-art attacks in terms of attack success rate and distance to the target model.
****

Generalization Error Bound for Hyperbolic Ordinal Embedding

Atsushi Suzuki, Atsushi Nitanda, Jing Wang, Linchuan Xu, Kenji Yamanishi, Marc Cavazza

Hyperbolic ordinal embedding (HOE) represents entities as points in hyperbolic space so that they agree as well as possible with given constraints in the form of entity i is more similar to entity j than to entity k. It has been experimentally shown that HOE can obtain representations of hierarchical data such as a knowledge base and a citation network effectively, owing to hyperbolic space’s exponential growth property. However, its theoretical analysis has been limited to ideal noiseless settings, and its generalization error in compensation for hyperbolic space’s exponential representation ability has not been guaranteed. The difficulty is that existing generalization error bound derivations for ordinal embedding based on the Gramian matrix are not applicable in HOE, since hyperbolic space is not inner-product space. In this paper, through our novel characterization of HOE with decomposed Lorentz Gramian matrices, we provide a generalization error bound of HOE for the first time, which is at most exponential with respect to the embedding space’s radius. Our comparison between the bounds of HOE and Euclidean ordinal embedding shows that HOE’s generalization error comes at a reasonable cost considering its exponential representation ability.
****

Of Moments and Matching: A Game-Theoretic Framework for Closing the Imitation Gap

Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, Steven Wu

We provide a unifying view of a large family of previous imitation learning algorithms through the lens of moment matching. At its core, our classification scheme is based on whether the learner attempts to match (1) reward or (2) action-value moments of the expert’s behavior, with each option leading to differing algorithmic approaches. By considering adversarially chosen divergences between learner and expert behavior, we are able to derive bounds on policy performance that apply for all algorithms in each of these classes, the first to our knowledge. We also introduce the notion of moment recoverability, implicit in many previous analyses of imitation learning, which allows us to cleanly delineate how well each algorithmic family is able to mitigate compounding errors. We derive three novel algorithm templates (AdVIL, AdRIL, and DAeQuIL) with strong guarantees, simple implementation, and competitive empirical performance.
****

Parallel tempering on optimized paths

Saifuddin Syed, Vittorio Romaniello, Trevor Campbell, Alexandre Bouchard-Cote

Parallel tempering (PT) is a class of Markov chain Monte Carlo algorithms that constructs a path of distributions annealing between a tractable reference and an

intractable target, and then interchanges states along the path to improve mixing in the target. The performance of PT depends on how quickly a sample from the reference distribution makes its way to the target, which in turn depends on the particular path of annealing distributions. However, past work on PT has used only simple paths constructed from convex combinations of the reference and target log-densities. This paper begins by demonstrating that this path performs poorly in the setting where the reference and target are nearly mutually singular. To address this issue, we expand the framework of PT to general families of paths, formulate the choice of path as an optimization problem that admits tractable gradient estimates, and propose a flexible new family of spline interpolation paths for use in practice. Theoretical and empirical results both demonstrate that our proposed methodology breaks previously-established upper performance limits for traditional paths.
****
Robust Representation Learning via Perceptual Similarity Metrics
Saeid A Taghanaki, Kristy Choi, Amir Hosein Khasahmadi, Anirudh Goyal
A fundamental challenge in artificial intelligence is learning useful representations of data that yield good performance on a downstream classification task, without overfitting to spurious input features. Extracting such task-relevant predictive information becomes particularly difficult for noisy and high-dimensional real-world data. In this work, we propose Contrastive Input Morphing (CIM), a representation learning framework that learns input-space transformations of the data to mitigate the effect of irrelevant input features on downstream performance. Our method leverages a perceptual similarity metric via a triplet loss to ensure that the transformation preserves task-relevant information. Empirically, we demonstrate the efficacy of our approach on various tasks which typically suffer from the presence of spurious correlations: classification with nuisance information, out-of-distribution generalization, and preservation of subgroup accuracies. We additionally show that CIM is complementary to other mutual information-based representation learning techniques, and demonstrate that it improves the performance of variational information bottleneck (VIB) when used in conjunction.
****
DriftSurf: Stable-State / Reactive-State Learning under Concept Drift
Ashraf Tahmasbi, Ellango Jothimurugesan, Srikanta Tirthapura, Phillip B Gibbons
When learning from streaming data, a change in the data distribution, also known as concept drift, can render a previously-learned model inaccurate and require training a new model. We present an adaptive learning algorithm that extends previous drift-detection-based methods by incorporating drift detection into a broader stable-state/reactive-state process. The advantage of our approach is that we can use aggressive drift detection in the stable state to achieve a high detection rate, but mitigate the false positive rate of standalone drift detection via a reactive state that reacts quickly to true drifts while eliminating most false positives. The algorithm is generic in its base learner and can be applied across a variety of supervised learning problems. Our theoretical analysis shows that the risk of the algorithm is (i) statistically better than standalone drift detection and (ii) competitive to an algorithm with oracle knowledge of when (abrupt) drifts occur. Experiments on synthetic and real datasets with concept drifts confirm our theoretical analysis.
****
Sinkhorn Label Allocation: Semi-Supervised Classification via Annealed Self-Training
Kai Sheng Tai, Peter D Bailis, Gregory Valiant
Self-training is a standard approach to semi-supervised learning where the learner’s own predictions on unlabeled data are used as supervision during training. In this paper, we reinterpret this label assignment process as an optimal transportation problem between examples and classes, wherein the cost of assigning an example to a class is mediated by the current predictions of the classifier. This formulation facilitates a practical annealing strategy for label assignment and allows for the inclusion of prior knowledge on class proportions via flexible

upper bound constraints. The solutions to these assignment problems can be effic iently approximated using Sinkhorn iteration, thus enabling their use in the inn er loop of standard stochastic optimization algorithms. We demonstrate the effec tiveness of our algorithm on the CIFAR-10, CIFAR-100, and SVHN datasets in compa rison with FixMatch, a state-of-the-art self-training algorithm.
****
Approximation Theory Based Methods for RKHS Bandits
Sho Takemori, Masahiro Sato
The RKHS bandit problem (also called kernelized multi-armed bandit problem) is a n online optimization problem of non-linear functions with noisy feedback. Altho ugh the problem has been extensively studied, there are unsatisfactory results f or some problems compared to the well-studied linear bandit case. Specifically, there is no general algorithm for the adversarial RKHS bandit problem. In additi on, high computational complexity of existing algorithms hinders practical appli cation. We address these issues by considering a novel amalgamation of approxima tion theory and the misspecified linear bandit problem. Using an approximation m ethod, we propose efficient algorithms for the stochastic RKHS bandit problem an d the first general algorithm for the adversarial RKHS bandit problem. Furthermo re, we empirically show that one of our proposed methods has comparable cumulati ve regret to IGP-UCB and its running time is much shorter.
****
Supervised Tree-Wasserstein Distance
Yuki Takezawa, Ryoma Sato, Makoto Yamada
To measure the similarity of documents, the Wasserstein distance is a powerful t ool, but it requires a high computational cost. Recently, for fast computation o f the Wasserstein distance, methods for approximating the Wasserstein distance u sing a tree metric have been proposed. These tree-based methods allow fast compa risons of a large number of documents; however, they are unsupervised and do not learn task-specific distances. In this work, we propose the Supervised Tree-Was serstein (STW) distance, a fast, supervised metric learning method based on the tree metric. Specifically, we rewrite the Wasserstein distance on the tree metri c by the parent-child relationships of a tree, and formulate it as a continuous optimization problem using a contrastive loss. Experimentally, we show that the STW distance can be computed fast, and improves the accuracy of document classif ication tasks. Furthermore, the STW distance is formulated by matrix multiplicat ions, runs on a GPU, and is suitable for batch processing. Therefore, we show th at the STW distance is extremely efficient when comparing a large number of docu ments.
****
EfficientNetV2: Smaller Models and Faster Training
Mingxing Tan, Quoc Le
This paper introduces EfficientNetV2, a new family of convolutional networks tha t have faster training speed and better parameter efficiency than previous model s. To develop these models, we use a combination of training-aware neural archit ecture search and scaling, to jointly optimize training speed and parameter effi ciency. The models were searched from the search space enriched with new ops suc h as Fused-MBConv. Our experiments show that EfficientNetV2 models train much fa ster than state-of-the-art models while being up to 6.8x smaller. Our training c an be further sped up by progressively increasing the image size during training , but it often causes a drop in accuracy. To compensate for this accuracy drop, we propose an improved method of progressive learning, which adaptively adjusts regularization (e.g. data augmentation) along with image size. With progressive learning, our EfficientNetV2 significantly outperforms previous models on ImageN et and CIFAR/Cars/Flowers datasets. By pretraining on the same ImageNet21k, our EfficientNetV2 achieves 87.3% top-1 accuracy on ImageNet ILSVRC2012, outperformi ng the recent ViT by 2.0% accuracy while training 5x-11x faster using the same c omputing resources.
****
SGA: A Robust Algorithm for Partial Recovery of Tree-Structured Graphical Models with Noisy Samples

Anshoo Tandon, Aldric Han, Vincent Tan
We consider learning Ising tree models when the observations from the nodes are corrupted by independent but non-identically distributed noise with unknown statistics. Katiyar et al. (2020) showed that although the exact tree structure cannot be recovered, one can recover a partial tree structure; that is, a structure belonging to the equivalence class containing the true tree. This paper presents a systematic improvement of Katiyar et al. (2020). First, we present a novel impossibility result by deriving a bound on the necessary number of samples for partial recovery. Second, we derive a significantly improved sample complexity result in which the dependence on the minimum correlation $\rho_{\min}$ is $\rho_{\min}^{-8}$ instead of $\rho_{\min}^{-24}$. Finally, we propose Symmetrized Geometric Averaging (SGA), a more statistically robust algorithm for partial tree recovery. We provide error exponent analyses and extensive numerical results on a variety of trees to show that the sample complexity of SGA is significantly better than the algorithm of Katiyar et al. (2020). SGA can be readily extended to Gaussian models and is shown via numerical experiments to be similarly superior.
****
1-bit Adam: Communication Efficient Large-Scale Training with Adam’s Convergence Speed
Hanlin Tang, Shaoduo Gan, Ammar Ahmad Awan, Samyam Rajbhandari, Conglong Li, Xiangru Lian, Ji Liu, Ce Zhang, Yuxiong He
Scalable training of large models (like BERT and GPT-3) requires careful optimization rooted in model design, architecture, and system capabilities. From a system standpoint, communication has become a major bottleneck, especially on commodity systems with standard TCP interconnects that offer limited network bandwidth. Communication compression is an important technique to reduce training time on such systems. One of the most effective ways to compress communication is via error compensation compression, which offers robust convergence speed, even under 1-bit compression. However, state-of-the-art error compensation techniques only work with basic optimizers like SGD and momentum SGD, which are linearly dependent on the gradients. They do not work with non-linear gradient-based optimizers like Adam, which offer state-of-the-art convergence efficiency and accuracy for models like BERT. In this paper, we propose 1-bit Adam that reduces the communication volume by up to 5x, offers much better scalability, and provides the same convergence speed as uncompressed Adam. Our key finding is that Adam’s variance becomes stable (after a warmup phase) and can be used as a fixed precondition for the rest of the training (compression phase). We performed experiments on up to 256 GPUs and show that 1-bit Adam enables up to 3.3x higher throughput for BERT-Large pre-training and up to 2.9x higher throughput for SQuAD fine-tuning. In addition, we provide theoretical analysis for 1-bit Adam.
****
Taylor Expansion of Discount Factors
Yunhao Tang, Mark Rowland, Remi Munos, Michal Valko
In practical reinforcement learning (RL), the discount factor used for estimating value functions often differs from that used for defining the evaluation objective. In this work, we study the effect that this discrepancy of discount factors has during learning, and discover a family of objectives that interpolate value functions of two distinct discount factors. Our analysis suggests new ways for estimating value functions and performing policy optimization updates, which demonstrate empirical performance gains. This framework also leads to new insights on commonly-used deep RL heuristic modifications to policy optimization algorithms.
****
REPAINT: Knowledge Transfer in Deep Reinforcement Learning
Yunzhe Tao, Sahika Genc, Jonathan Chung, Tao Sun, Sunil Mallya
Accelerating learning processes for complex tasks by leveraging previously learned tasks has been one of the most challenging problems in reinforcement learning, especially when the similarity between source and target tasks is low. This work proposes REPresentation And INstance Transfer (REPAINT) algorithm for knowledge transfer in deep reinforcement learning. REPAINT not only transfers the repre

sentation of a pre-trained teacher policy in the on-policy learning, but also uses an advantage-based experience selection approach to transfer useful samples collected following the teacher policy in the off-policy learning. Our experimental results on several benchmark tasks show that REPAINT significantly reduces the total training time in generic cases of task similarity. In particular, when the source tasks are dissimilar to, or sub-tasks of, the target tasks, REPAINT outperforms other baselines in both training-time reduction and asymptotic performance of return scores.
****
Understanding the Dynamics of Gradient Flow in Overparameterized Linear models
Salma Tarmoun, Guilherme Franca, Benjamin D Haeffele, Rene Vidal
We provide a detailed analysis of the dynamics ofthe gradient flow in overparameterized two-layerlinear models. A particularly interesting featureof this model is that its nonlinear dynamics can beexactly solved as a consequence of a large num-ber of conservation laws that constrain the systemto follow particular trajectories. More precisely,the gradient flow preserves the difference of theGramian matrices of the input and output weights,and its convergence to equilibrium depends onboth the magnitude of that difference (which isfixed at initialization) and the spectrum of the data.In addition, and generalizing prior work, we proveour results without assuming small, balanced orspectral initialization for the weights. Moreover,we establish interesting mathematical connectionsbetween matrix factorization problems and differ-ential equations of the Riccati type.
****
Sequential Domain Adaptation by Synthesizing Distributionally Robust Experts
Bahar Taskesen, Man-Chung Yue, Jose Blanchet, Daniel Kuhn, Viet Anh Nguyen
Least squares estimators, when trained on few target domain samples, may predict poorly. Supervised domain adaptation aims to improve the predictive accuracy by exploiting additional labeled training samples from a source distribution that is close to the target distribution. Given available data, we investigate novel strategies to synthesize a family of least squares estimator experts that are robust with regard to moment conditions. When these moment conditions are specified using Kullback-Leibler or Wasserstein-type divergences, we can find the robust estimators efficiently using convex optimization. We use the Bernstein online aggregation algorithm on the proposed family of robust experts to generate predictions for the sequential stream of target test samples. Numerical experiments on real data show that the robust strategies systematically outperform non-robust interpolations of the empirical least squares estimators.
****
A Language for Counterfactual Generative Models
Zenna Tavares, James Koppel, Xin Zhang, Ria Das, Armando Solar-Lezama
We present Omega, a probabilistic programming language with support for counterfactual inference. Counterfactual inference means to observe some fact in the present, and infer what would have happened had some past intervention been taken, e.g. “given that medication was not effective at dose x, what is the probability that it would have been effective at dose 2x?.” We accomplish this by introducing a new operator to probabilistic programming akin to Pearl’s do, define its formal semantics, provide an implementation, and demonstrate its utility through examples in a variety of simulation models.
****
Synthesizer: Rethinking Self-Attention for Transformer Models
Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, Che Zheng
The dot product self-attention is known to be central and indispensable to state-of-the-art Transformer models. But is it really required? This paper investigates the true importance and contribution of the dot product-based self-attention mechanism on the performance of Transformer models. Via extensive experiments, we find that (1) random alignment matrices surprisingly perform quite competitively and (2) learning attention weights from token-token (query-key) interactions is useful but not that important after all. To this end, we propose \textsc{Synthesizer}, a model that learns synthetic attention weights without token-token interactions. In our experiments, we first show that simple Synthesizers achieve h

ighly competitive performance when compared against vanilla Transformer models across a range of tasks, including machine translation, language modeling, text generation and GLUE/SuperGLUE benchmarks. When composed with dot product attention, we find that Synthesizers consistently outperform Transformers. Moreover, we conduct additional comparisons of Synthesizers against Dynamic Convolutions, showing that simple Random Synthesizer is not only $60%$ faster but also improves perplexity by a relative $3.5%$. Finally, we show that simple factorized Synthesizers can outperform Linformers on encoding only tasks.
****
OmniNet: Omnidirectional Representations from Transformers
Yi Tay, Mostafa Dehghani, Vamsi Aribandi, Jai Gupta, Philip M Pham, Zhen Qin, Dara Bahri, Da-Cheng Juan, Donald Metzler
This paper proposes Omnidirectional Representations from Transformers (OMNINET). In OmniNet, instead of maintaining a strictly horizon-tal receptive field, each token is allowed to attend to all tokens in the entire network. This process can also be interpreted as a form of extreme or intensive attention mechanism that has the receptive field of the entire width and depth of the network. To this end, the omnidirectional attention is learned via a meta-learner, which is essentially another self-attention based model. In order to mitigate the computationally expensive costs of full receptive field attention, we leverage efficient self-attention models such as kernel-based, low-rank attention and/or Big Bird as the meta-learner. Extensive experiments are conducted on autoregressive language modeling(LM1B, C4), Machine Translation, Long Range Arena (LRA), and Image Recognition.The experiments show that OmniNet achieves considerable improvements across these tasks, including achieving state-of-the-art performance on LM1B,WMT’14 En-De/En-Fr, and Long Range Arena.Moreover, using omnidirectional representation in Vision Transformers leads to significant improvements on image recognition tasks on both few-shot learning and fine-tuning setups.
****
T-SCI: A Two-Stage Conformal Inference Algorithm with Guaranteed Coverage for Cox-MLP
Jiaye Teng, Zeren Tan, Yang Yuan
It is challenging to deal with censored data, where we only have access to the incomplete information of survival time instead of its exact value. Fortunately, under linear predictor assumption, people can obtain guaranteed coverage for the confidence interval of survival time using methods like Cox Regression. However, when relaxing the linear assumption with neural networks (e.g., Cox-MLP \citep{katzman2018deepsurv,kvamme2019time}), we lose the guaranteed coverage. To recover the guaranteed coverage without linear assumption, we propose two algorithms based on conformal inference. In the first algorithm \emph{WCCI}, we revisit weighted conformal inference and introduce a new non-conformity score based on partial likelihood. We then propose a two-stage algorithm \emph{T-SCI}, where we run WCCI in the first stage and apply quantile conformal inference to calibrate the results in the second stage. Theoretical analysis shows that T-SCI returns guaranteed coverage under milder assumptions than WCCI. We conduct extensive experiments on synthetic data and real data using different methods, which validate our analysis.
****
Moreau-Yosida f-divergences
Dávid Terjék
Variational representations of f-divergences are central to many machine learning algorithms, with Lipschitz constrained variants recently gaining attention. Inspired by this, we define the Moreau-Yosida approximation of f-divergences with respect to the Wasserstein-1 metric. The corresponding variational formulas provide a generalization of a number of recent results, novel special cases of interest and a relaxation of the hard Lipschitz constraint. Additionally, we prove that the so-called tight variational representation of f-divergences can be to be taken over the quotient space of Lipschitz functions, and give a characterization of functions achieving the supremum in the variational representation. On the practical side, we propose an algorithm to calculate the tight convex co

njugate of f-divergences compatible with automatic differentiation frameworks. As an application of our results, we propose the Moreau-Yosida f-GAN, providing an implementation of the variational formulas for the Kullback-Leibler, reverse Kullback-Leibler, χ^2, reverse χ^2, squared Hellinger, Jensen-Shannon, Jeffreys, triangular discrimination and total variation divergences as GANs trained on CIFAR-10, leading to competitive results and a simple solution to the problem of uniqueness of the optimal critic.
****
Understanding Invariance via Feedforward Inversion of Discriminatively Trained Classifiers
Piotr Teterwak, Chiyuan Zhang, Dilip Krishnan, Michael C Mozer
A discriminatively trained neural net classifier can fit the training data perfectly if all information about its input other than class membership has been discarded prior to the output layer. Surprisingly, past research has discovered that some extraneous visual detail remains in the unnormalized logits. This finding is based on inversion techniques that map deep embeddings back to images. We explore this phenomenon further using a novel synthesis of methods, yielding a feedforward inversion model that produces remarkably high fidelity reconstructions, qualitatively superior to those of past efforts. When applied to an adversarially robust classifier model, the reconstructions contain sufficient local detail and global structure that they might be confused with the original image in a quick glance, and the object category can clearly be gleaned from the reconstruction. Our approach is based on BigGAN (Brock, 2019), with conditioning on logits instead of one-hot class labels. We use our reconstruction model as a tool for exploring the nature of representations, including: the influence of model architecture and training objectives (specifically robust losses), the forms of invariance that networks achieve, representational differences between correctly and incorrectly classified images, and the effects of manipulating logits and images. We believe that our method can inspire future investigations into the nature of information flow in a neural net and can provide diagnostics for improving discriminative models. We provide pre-trained models and visualizations at \url{https://sites.google.com/view/understanding-invariance/home}.
****
Resource Allocation in Multi-armed Bandit Exploration: Overcoming Sublinear Scaling with Adaptive Parallelism
Brijen Thananjeyan, Kirthevasan Kandasamy, Ion Stoica, Michael Jordan, Ken Goldberg, Joseph Gonzalez
We study exploration in stochastic multi-armed bandits when we have access to a divisible resource that can be allocated in varying amounts to arm pulls. We focus in particular on the allocation of distributed computing resources, where we may obtain results faster by allocating more resources per pull, but might have reduced throughput due to nonlinear scaling. For example, in simulation-based scientific studies, an expensive simulation can be sped up by running it on multiple cores. This speed-up however, is partly offset by the communication among cores, which results in lower throughput than if fewer cores were allocated to run more trials in parallel. In this paper, we explore these trade-offs in two settings. First, in a fixed confidence setting, we need to find the best arm with a given target success probability as quickly as possible. We propose an algorithm which trades off between information accumulation and throughput and show that the time taken can be upper bounded by the solution of a dynamic program whose inputs are the gaps between the sub-optimal and optimal arms. We also prove a matching hardness result. Second, we present an algorithm for a fixed deadline setting, where we are given a time deadline and need to maximize the probability of finding the best arm. We corroborate our theoretical insights with simulation experiments that show that the algorithms consistently match or outperform baseline algorithms on a variety of problem instances.
****
Monte Carlo Variational Auto-Encoders
Achille Thin, Nikita Kotelevskii, Arnaud Doucet, Alain Durmus, Eric Moulines, Maxim Panov

Variational auto-encoders (VAE) are popular deep latent variable models which ar
e trained by maximizing an Evidence Lower Bound (ELBO). To obtain tighter ELBO a
nd hence better variational approximations, it has been proposed to use importan
ce sampling to get a lower variance estimate of the evidence. However, importanc
e sampling is known to perform poorly in high dimensions. While it has been sugg
ested many times in the literature to use more sophisticated algorithms such as
Annealed Importance Sampling (AIS) and its Sequential Importance Sampling (SIS)
extensions, the potential benefits brought by these advanced techniques have nev
er been realized for VAE: the AIS estimate cannot be easily differentiated, whil
e SIS requires the specification of carefully chosen backward Markov kernels. In
this paper, we address both issues and demonstrate the performance of the resul
ting Monte Carlo VAEs on a variety of applications.
****
Efficient Generative Modelling of Protein Structure Fragments using a Deep Marko
v Model
Christian B Thygesen, Christian Skjødt Steenmans, Ahmad Salim Al-Sibahi, Lys San
z Moreta, Anders Bundgård Sørensen, Thomas Hamelryck
Fragment libraries are often used in protein structure prediction, simulation an
d design as a means to significantly reduce the vast conformational search space
. Current state-of-the-art methods for fragment library generation do not proper
ly account for aleatory and epistemic uncertainty, respectively due to the dynam
ic nature of proteins and experimental errors in protein structures. Additionall
y, they typically rely on information that is not generally or readily available
, such as homologous sequences, related protein structures and other complementa
ry information. To address these issues, we developed BIFROST, a novel take on t
he fragment library problem based on a Deep Markov Model architecture combined w
ith directional statistics for angular degrees of freedom, implemented in the de
ep probabilistic programming language Pyro. BIFROST is a probabilistic, generati
ve model of the protein backbone dihedral angles conditioned solely on the amino
acid sequence. BIFROST generates fragment libraries with a quality on par with
current state-of-the-art methods at a fraction of the run-time, while requiring
considerably less information and allowing efficient evaluation of probabilities
.
****
Understanding self-supervised learning dynamics without contrastive pairs
Yuandong Tian, Xinlei Chen, Surya Ganguli
While contrastive approaches of self-supervised learning (SSL) learn representat
ions by minimizing the distance between two augmented views of the same data poi
nt (positive pairs) and maximizing views from different data points (negative pa
irs), recent \emph{non-contrastive} SSL (e.g., BYOL and SimSiam) show remarkable
performance {\it without} negative pairs, with an extra learnable predictor and
a stop-gradient operation. A fundamental question rises: why they do not collap
se into trivial representation? In this paper, we answer this question via a sim
ple theoretical study and propose a novel approach, \ourmethod{}, that \emph{dir
ectly} sets the linear predictor based on the statistics of its inputs, rather t
han trained with gradient update. On ImageNet, it performs comparably with more
complex two-layer non-linear predictors that employ BatchNorm and outperforms li
near predictor by $2.5%$ in 300-epoch training (and $5%$ in 60-epoch). \ourmetho
d{} is motivated by our theoretical study of the nonlinear learning dynamics of
non-contrastive SSL in simple linear networks. Our study yields conceptual insig
hts into how non-contrastive SSL methods learn, how they avoid representational
collapse, and how multiple factors, like predictor networks, stop-gradients, exp
onential moving averages, and weight decay all come into play. Our simple theory
recapitulates the results of real-world ablation studies in both STL-10 and Ima
geNet. Code is released\footnote{\url{https://github.com/facebookresearch/luckma
tters/tree/master/ssl}}.
****
Online Learning in Unknown Markov Games
Yi Tian, Yuanhao Wang, Tiancheng Yu, Suvrit Sra
We study online learning in unknown Markov games, a problem that arises in episo

dic multi-agent reinforcement learning where the actions of the opponents are un observable. We show that in this challenging setting, achieving sublinear regret against the best response in hindsight is statistically hard. We then consider a weaker notion of regret by competing with the \emph{minimax value} of the game , and present an algorithm that achieves a sublinear $\tilde{\mathcal{O}}(K^{2/3})$ regret after K episodes. This is the first sublinear regret bound (to our knowledge) for online learning in unknown Markov games. Importantly, our regret bound is independent of the size of the opponents' action spaces. As a result, even when the opponents' actions are fully observable, our regret bound improves upon existing analysis (e.g., (Xie et al., 2020)) by an exponential factor in the number of opponents.
****
BORE: Bayesian Optimization by Density-Ratio Estimation
Louis C Tiao, Aaron Klein, Matthias W Seeger, Edwin V. Bonilla, Cedric Archambeau, Fabio Ramos
Bayesian optimization (BO) is among the most effective and widely-used blackbox optimization methods. BO proposes solutions according to an explore-exploit trade-off criterion encoded in an acquisition function, many of which are computed from the posterior predictive of a probabilistic surrogate model. Prevalent among these is the expected improvement (EI). The need to ensure analytical tractability of the predictive often poses limitations that can hinder the efficiency and applicability of BO. In this paper, we cast the computation of EI as a binary classification problem, building on the link between class-probability estimation and density-ratio estimation, and the lesser-known link between density-ratios and EI. By circumventing the tractability constraints, this reformulation provides numerous advantages, not least in terms of expressiveness, versatility, and scalability.
****
Nonparametric Decomposition of Sparse Tensors
Conor Tillinghast, Shandian Zhe
Tensor decomposition is a powerful framework for multiway data analysis. Despite the success of existing approaches, they ignore the sparse nature of the tensor data in many real-world applications, explicitly or implicitly assuming dense tensors. To address this model misspecification and to exploit the sparse tensor structures, we propose Nonparametric dEcomposition of Sparse Tensors (\ours), which can capture both the sparse structure properties and complex relationships between the tensor nodes to enhance the embedding estimation. Specifically, we first use completely random measures to construct tensor-valued random processes. We prove that the entry growth is much slower than that of the corresponding tensor size, which implies sparsity. Given finite observations (\ie projections), we then propose two nonparametric decomposition models that couple Dirichlet processes and Gaussian processes to jointly sample the sparse entry indices and the entry values (the latter as a nonlinear mapping of the embeddings), so as to encode both the structure properties and nonlinear relationships of the tensor nodes into the embeddings. Finally, we use the stick-breaking construction and random Fourier features to develop a scalable, stochastic variational learning algorithm. We show the advantage of our approach in sparse tensor generation, and entry index and value prediction in several real-world applications.
****
Probabilistic Programs with Stochastic Conditioning
David Tolpin, Yuan Zhou, Tom Rainforth, Hongseok Yang
We tackle the problem of conditioning probabilistic programs on distributions of observable variables. Probabilistic programs are usually conditioned on samples from the joint data distribution, which we refer to as deterministic conditioning. However, in many real-life scenarios, the observations are given as marginal distributions, summary statistics, or samplers. Conventional probabilistic programming systems lack adequate means for modeling and inference in such scenarios. We propose a generalization of deterministic conditioning to stochastic conditioning, that is, conditioning on the marginal distribution of a variable taking a particular form. To this end, we first define the formal notion of stochastic

conditioning and discuss its key properties. We then show how to perform inferen ce in the presence of stochastic conditioning. We demonstrate potential usage of stochastic conditioning on several case studies which involve various kinds of stochastic conditioning and are difficult to solve otherwise. Although we presen t stochastic conditioning in the context of probabilistic programming, our forma lization is general and applicable to other settings.
****
Deep Continuous Networks
Nergis Tomen, Silvia-Laura Pintea, Jan Van Gemert
CNNs and computational models of biological vision share some fundamental princi ples, which opened new avenues of research. However, fruitful cross-field resear ch is hampered by conventional CNN architectures being based on spatially and de pthwise discrete representations, which cannot accommodate certain aspects of bi ological complexity such as continuously varying receptive field sizes and dynam ics of neuronal responses. Here we propose deep continuous networks (DCNs), whic h combine spatially continuous filters, with the continuous depth framework of n eural ODEs. This allows us to learn the spatial support of the filters during tr aining, as well as model the continuous evolution of feature maps, linking DCNs closely to biological models. We show that DCNs are versatile and highly applica ble to standard image classification and reconstruction problems, where they imp rove parameter and data efficiency, and allow for meta-parametrization. We illus trate the biological plausibility of the scale distributions learned by DCNs and explore their performance in a neuroscientifically inspired pattern completion task. Finally, we investigate an efficient implementation of DCNs by changing in put contrast.
****
Diffusion Earth Mover’s Distance and Distribution Embeddings
Alexander Y Tong, Guillaume Huguet, Amine Natik, Kincaid Macdonald, Manik Kuchro o, Ronald Coifman, Guy Wolf, Smita Krishnaswamy
We propose a new fast method of measuring distances between large numbers of rel ated high dimensional datasets called the Diffusion Earth Mover’s Distance (EMD) . We model the datasets as distributions supported on common data graph that is derived from the affinity matrix computed on the combined data. In such cases wh ere the graph is a discretization of an underlying Riemannian closed manifold, w e prove that Diffusion EMD is topologically equivalent to the standard EMD with a geodesic ground distance. Diffusion EMD can be computed in {Õ}(n) time and is more accurate than similarly fast algorithms such as tree-based EMDs. We also sh ow Diffusion EMD is fully differentiable, making it amenable to future uses in g radient-descent frameworks such as deep neural networks. Finally, we demonstrate an application of Diffusion EMD to single cell data collected from 210 COVID-19 patient samples at Yale New Haven Hospital. Here, Diffusion EMD can derive dist ances between patients on the manifold of cells at least two orders of magnitude faster than equally accurate methods. This distance matrix between patients can be embedded into a higher level patient manifold which uncovers structure and h eterogeneity in patients. More generally, Diffusion EMD is applicable to all dat asets that are massively collected in parallel in many medical and biological sy stems.
****
Training data-efficient image transformers & distillation through attention
Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayro lles, Herve Jegou
Recently, neural networks purely based on attention were shown to address image understanding tasks such as image classification. These high-performing vision t ransformers are pre-trained with hundreds of millions of images using a large in frastructure, thereby limiting their adoption. In this work, we produce competit ive convolution-free transformers trained on ImageNet only using a single comput er in less than 3 days. Our reference vision transformer (86M parameters) achiev es top-1 accuracy of 83.1% (single-crop) on ImageNet with no external data. We a lso introduce a teacher-student strategy specific to transformers. It relies on a distillation token ensuring that the student learns from the teacher through a

ttention, typically from a convnet teacher. The learned transformers are competi
tive (85.2% top-1 acc.) with the state of the art on ImageNet, and similarly whe
n transferred to other tasks. We will share our code and models.
****
Conservative Objective Models for Effective Offline Model-Based Optimization
Brandon Trabucco, Aviral Kumar, Xinyang Geng, Sergey Levine
In this paper, we aim to solve data-driven model-based optimization (MBO) proble
ms, where the goal is to find a design input that maximizes an unknown objective
 function provided access to only a static dataset of inputs and their correspon
ding objective values. Such data-driven optimization procedures are the only pra
ctical methods in many real-world domains where active data collection is expens
ive (e.g., when optimizing over proteins) or dangerous (e.g., when optimizing ov
er aircraft designs, actively evaluating malformed aircraft designs is unsafe).
Typical methods for MBO that optimize the input against a learned model of the u
nknown score function are affected by erroneous overestimation in the learned mo
del caused due to distributional shift, that drives the optimizer to low-scoring
 or invalid inputs. To overcome this, we propose conservative objective models (
COMs), a method that learns a model of the objective function which lower bounds
 the actual value of the ground-truth objective on out-of-distribution inputs an
d uses it for optimization. In practice, COMs outperform a number existing metho
ds on a wide range of MBO problems, including optimizing controller parameters,
robot morphologies, and superconducting materials.
****
Sparse within Sparse Gaussian Processes using Neighbor Information
Gia-Lac Tran, Dimitrios Milios, Pietro Michiardi, Maurizio Filippone
Approximations to Gaussian processes (GPs) based on inducing variables, combined
 with variational inference techniques, enable state-of-the-art sparse approache
s to infer GPs at scale through mini-batch based learning. In this work, we furt
her push the limits of scalability of sparse GPs by allowing large number of ind
ucing variables without imposing a special structure on the inducing inputs. In
particular, we introduce a novel hierarchical prior, which imposes sparsity on t
he set of inducing variables. We treat our model variationally, and we experimen
tally show considerable computational gains compared to standard sparse GPs when
 sparsity on the inducing variables is realized considering the nearest inducing
 inputs of a random mini-batch of the data. We perform an extensive experimental
 validation that demonstrates the effectiveness of our approach compared to the
state-of-the-art. Our approach enables the possibility to use sparse GPs using a
 large number of inducing points without incurring a prohibitive computational c
ost.
****
SMG: A Shuffling Gradient-Based Method with Momentum
Trang H Tran, Lam M Nguyen, Quoc Tran-Dinh
We combine two advanced ideas widely used in optimization for machine learning:
\textit{shuffling} strategy and \textit{momentum} technique to develop a novel s
huffling gradient-based method with momentum, coined \textbf{S}huffling \textbf{
M}omentum \textbf{G}radient (SMG), for non-convex finite-sum optimization proble
ms. While our method is inspired by momentum techniques, its update is fundament
ally different from existing momentum-based methods. We establish state-of-the-a
rt convergence rates of SMG for any shuffling strategy using either constant or
diminishing learning rate under standard assumptions (i.e. \textit{L-smoothnes
s} and \textit{bounded variance}). When the shuffling strategy is fixed, we deve
lop another new algorithm that is similar to existing momentum methods, and prov
e the same convergence rates for this algorithm under the L-smoothness and bou
nded gradient assumptions. We demonstrate our algorithms via numerical simulatio
ns on standard datasets and compare them with existing shuffling methods. Our te
sts have shown encouraging performance of the new algorithms.
****
Bayesian Optimistic Optimisation with Exponentially Decaying Regret
Hung Tran-The, Sunil Gupta, Santu Rana, Svetha Venkatesh
Bayesian optimisation (BO) is a well known algorithm for finding the global opti

mum of expensive, black-box functions. The current practical BO algorithms have
regret bounds ranging from $\mathcal{O}(\frac{logN}{\sqrt{N}})$ to $\mathcal O(e
^{-\sqrt{N}})$, where N is the number of evaluations. This paper explores the
possibility of improving the regret bound in the noise-free setting by intertwin
ing concepts from BO and optimistic optimisation methods which are based on part
itioning the search space. We propose the BOO algorithm, a first practical appro
ach which can achieve an exponential regret bound with order $\mathcal O(N^{-\sq
rt{N}})$ under the assumption that the objective function is sampled from a Gaus
sian process with a Matérn kernel with smoothness parameter $\nu > 4 +\frac{D}{2
}$, where D is the number of dimensions. We perform experiments on optimisatio
n of various synthetic functions and machine learning hyperparameter tuning task
s and show that our algorithm outperforms baselines.
****
On Disentangled Representations Learned from Correlated Data
Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Di
ttadi, Anirudh Goyal, Bernhard Schölkopf, Stefan Bauer
The focus of disentanglement approaches has been on identifying independent fact
ors of variation in data. However, the causal variables underlying real-world ob
servations are often not statistically independent. In this work, we bridge the
gap to real-world scenarios by analyzing the behavior of the most prominent dise
ntanglement approaches on correlated data in a large-scale empirical study (incl
uding 4260 models). We show and quantify that systematically induced correlation
s in the dataset are being learned and reflected in the latent representations,
which has implications for downstream applications of disentanglement such as fa
irness. We also demonstrate how to resolve these latent correlations, either usi
ng weak supervision during training or by post-hoc correcting a pre-trained mode
l with a small number of labels.
****
A New Formalism, Method and Open Issues for Zero-Shot Coordination
Johannes Treutlein, Michael Dennis, Caspar Oesterheld, Jakob Foerster
In many coordination problems, independently reasoning humans are able to discov
er mutually compatible policies. In contrast, independently trained self-play po
licies are often mutually incompatible. Zero-shot coordination (ZSC) has recentl
y been proposed as a new frontier in multi-agent reinforcement learning to addre
ss this fundamental issue. Prior work approaches the ZSC problem by assuming pla
yers can agree on a shared learning algorithm but not on labels for actions and
observations, and proposes other-play as an optimal solution. However, until now
, this “label-free” problem has only been informally defined. We formalize this
setting as the label-free coordination (LFC) problem by defining the label-free
coordination game. We show that other-play is not an optimal solution to the LFC
 problem as it fails to consistently break ties between incompatible maximizers
of the other-play objective. We introduce an extension of the algorithm, other-p
lay with tie-breaking, and prove that it is optimal in the LFC problem and an eq
uilibrium in the LFC game. Since arbitrary tie-breaking is precisely what the ZS
C setting aims to prevent, we conclude that the LFC problem does not reflect the
 aims of ZSC. To address this, we introduce an alternative informal operationali
zation of ZSC as a starting point for future work.
****
Learning a Universal Template for Few-shot Dataset Generalization
Eleni Triantafillou, Hugo Larochelle, Richard Zemel, Vincent Dumoulin
Few-shot dataset generalization is a challenging variant of the well-studied few
-shot classification problem where a diverse training set of several datasets is
 given, for the purpose of training an adaptable model that can then learn class
es from \emph{new datasets} using only a few examples. To this end, we propose t
o utilize the diverse training set to construct a \emph{universal template}: a p
artial model that can define a wide array of dataset-specialized models, by plug
ging in appropriate components. For each new few-shot classification problem, ou
r approach therefore only requires inferring a small number of parameters to ins
ert into the universal template. We design a separate network that produces an i
nitialization of those parameters for each given task, and we then fine-tune its

proposed initialization via a few steps of gradient descent. Our approach is more parameter-efficient, scalable and adaptable compared to previous methods, and achieves the state-of-the-art on the challenging Meta-Dataset benchmark.
****

Provable Meta-Learning of Linear Representations

Nilesh Tripuraneni, Chi Jin, Michael Jordan

Meta-learning, or learning-to-learn, seeks to design algorithms that can utilize previous experience to rapidly learn new skills or adapt to new environments. Representation learning—a key tool for performing meta-learning—learns a data representation that can transfer knowledge across multiple tasks, which is essential in regimes where data is scarce. Despite a recent surge of interest in the practice of meta-learning, the theoretical underpinnings of meta-learning algorithms are lacking, especially in the context of learning transferable representations. In this paper, we focus on the problem of multi-task linear regression—in which multiple linear regression models share a common, low-dimensional linear representation. Here, we provide provably fast, sample-efficient algorithms to address the dual challenges of (1) learning a common set of features from multiple, related tasks, and (2) transferring this knowledge to new, unseen tasks. Both are central to the general problem of meta-learning. Finally, we complement these results by providing information-theoretic lower bounds on the sample complexity of learning these linear features.
****

Cumulants of Hawkes Processes are Robust to Observation Noise

William Trouleau, Jalal Etesami, Matthias Grossglauser, Negar Kiyavash, Patrick Thiran

Multivariate Hawkes processes (MHPs) are widely used in a variety of fields to model the occurrence of causally related discrete events in continuous time. Most state-of-the-art approaches address the problem of learning MHPs from perfect traces without noise. In practice, the process through which events are collected might introduce noise in the timestamps. In this work, we address the problem of learning the causal structure of MHPs when the observed timestamps of events are subject to random and unknown shifts, also known as random translations. We prove that the cumulants of MHPs are invariant to random translations, and therefore can be used to learn their underlying causal structure. Furthermore, we empirically characterize the effect of random translations on state-of-the-art learning methods. We show that maximum likelihood-based estimators are brittle, while cumulant-based estimators remain stable even in the presence of significant time shifts.
****

PixelTransformer: Sample Conditioned Signal Generation

Shubham Tulsiani, Abhinav Gupta

We propose a generative model that can infer a distribution for the underlying spatial signal conditioned on sparse samples e.g. plausible images given a few observed pixels. In contrast to sequential autoregressive generative models, our model allows conditioning on arbitrary samples and can answer distributional queries for any location. We empirically validate our approach across three image datasets and show that we learn to generate diverse and meaningful samples, with the distribution variance reducing given more observed pixels. We also show that our approach is applicable beyond images and can allow generating other types of spatial outputs e.g. polynomials, 3D shapes, and videos.
****

A Framework for Private Matrix Analysis in Sliding Window Model

Jalaj Upadhyay, Sarvagya Upadhyay

We perform a rigorous study of private matrix analysis when only the last W updates to matrices are considered useful for analysis. We show the existing framework in the non-private setting is not robust to noise required for privacy. We then propose a framework robust to noise and use it to give first efficient $o(W)$ space differentially private algorithms for spectral approximation, principal component analysis (PCA), multi-response linear regression, sparse PCA, and non-negative PCA. Prior to our work, no such result was known for sparse and non-ne

gative differentially private PCA even in the static data setting. We also give a lower bound to demonstrate the cost of privacy in the sliding window model.
****

Fast Projection Onto Convex Smooth Constraints
Ilnura Usmanova, Maryam Kamgarpour, Andreas Krause, Kfir Levy
The Euclidean projection onto a convex set is an important problem that arises in numerous constrained optimization tasks. Unfortunately, in many cases, computing projections is computationally demanding. In this work, we focus on projection problems where the constraints are smooth and the number of constraints is significantly smaller than the dimension. The runtime of existing approaches to solving such problems is either cubic in the dimension or polynomial in the inverse of the target accuracy. Conversely, we propose a simple and efficient primal-dual approach, with a runtime that scales only linearly with the dimension, and only logarithmically in the inverse of the target accuracy. We empirically demonstrate its performance, and compare it with standard baselines.
****

SGLB: Stochastic Gradient Langevin Boosting
Aleksei Ustimenko, Liudmila Prokhorenkova
This paper introduces Stochastic Gradient Langevin Boosting (SGLB) - a powerful and efficient machine learning framework that may deal with a wide range of loss functions and has provable generalization guarantees. The method is based on a special form of the Langevin diffusion equation specifically designed for gradient boosting. This allows us to theoretically guarantee the global convergence even for multimodal loss functions, while standard gradient boosting algorithms can guarantee only local optimum. We also empirically show that SGLB outperforms classic gradient boosting when applied to classification tasks with 0-1 loss function, which is known to be multimodal.
****

LTL2Action: Generalizing LTL Instructions for Multi-Task RL
Pashootan Vaezipoor, Andrew C Li, Rodrigo A Toro Icarte, Sheila A. Mcilraith
We address the problem of teaching a deep reinforcement learning (RL) agent to follow instructions in multi-task environments. Instructions are expressed in a well-known formal language {–} linear temporal logic (LTL) {–} and can specify a diversity of complex, temporally extended behaviours, including conditionals and alternative realizations. Our proposed learning approach exploits the compositional syntax and the semantics of LTL, enabling our RL agent to learn task-conditioned policies that generalize to new instructions, not observed during training. To reduce the overhead of learning LTL semantics, we introduce an environment-agnostic LTL pretraining scheme which improves sample-efficiency in downstream environments. Experiments on discrete and continuous domains target combinatorial task sets of up to $\sim10^{39}$ unique tasks and demonstrate the strength of our approach in learning to solve (unseen) tasks, given LTL instructions.
****

Active Deep Probabilistic Subsampling
Hans Van Gorp, Iris Huijben, Bastiaan S Veeling, Nicola Pezzotti, Ruud J. G. Van Sloun
Subsampling a signal of interest can reduce costly data transfer, battery drain, radiation exposure and acquisition time in a wide range of problems. The recently proposed Deep Probabilistic Subsampling (DPS) method effectively integrates subsampling in an end-to-end deep learning model, but learns a static pattern for all datapoints. We generalize DPS to a sequential method that actively picks the next sample based on the information acquired so far; dubbed Active-DPS (A-DPS). We validate that A-DPS improves over DPS for MNIST classification at high subsampling rates. Moreover, we demonstrate strong performance in active acquisition Magnetic Resonance Image (MRI) reconstruction, outperforming DPS and other deep learning methods.
****

CURI: A Benchmark for Productive Concept Learning Under Uncertainty
Ramakrishna Vedantam, Arthur Szlam, Maximillian Nickel, Ari Morcos, Brenden M Lake

Humans can learn and reason under substantial uncertainty in a space of infinitely many compositional, productive concepts. For example, if a scene with two blue spheres qualifies as “daxy,” one can reason that the underlying concept may require scenes to have “only blue spheres” or “only spheres” or “only two objects.” In contrast, standard benchmarks for compositional reasoning do not explicitly capture a notion of reasoning under uncertainty or evaluate compositional concept acquisition. We introduce a new benchmark, Compositional Reasoning Under Uncertainty (CURI) that instantiates a series of few-shot, meta-learning tasks in a productive concept space to evaluate different aspects of systematic generalization under uncertainty, including splits that test abstract understandings of disentangling, productive generalization, learning boolean operations, variable binding, etc. Importantly, we also contribute a model-independent “compositionality gap” to evaluate the difficulty of generalizing out-of-distribution along each of these axes, allowing objective comparison of the difficulty of each compositional split. Evaluations across a range of modeling choices and splits reveal substantial room for improvement on the proposed benchmark.
****
Towards Domain-Agnostic Contrastive Learning
Vikas Verma, Thang Luong, Kenji Kawaguchi, Hieu Pham, Quoc Le
Despite recent successes, most contrastive self-supervised learning methods are domain-specific, relying heavily on data augmentation techniques that require knowledge about a particular domain, such as image cropping and rotation. To overcome such limitation, we propose a domain-agnostic approach to contrastive learning, named DACL, that is applicable to problems where domain-specific data augmentations are not readily available. Key to our approach is the use of Mixup noise to create similar and dissimilar examples by mixing data samples differently either at the input or hidden-state levels. We theoretically analyze our method and show advantages over the Gaussian-noise based contrastive learning approach. To demonstrate the effectiveness of DACL, we conduct experiments across various domains such as tabular data, images, and graphs. Our results show that DACL not only outperforms other domain-agnostic noising methods, such as Gaussian-noise, but also combines well with domain-specific methods, such as SimCLR, to improve self-supervised visual representation learning.
****
Sparsifying Networks via Subdifferential Inclusion
Sagar Verma, Jean-Christophe Pesquet
Sparsifying deep neural networks is of paramount interest in many areas, especially when those networks have to be implemented on low-memory devices. In this article, we propose a new formulation of the problem of generating sparse weights for a pre-trained neural network. By leveraging the properties of standard nonlinear activation functions, we show that the problem is equivalent to an approximate subdifferential inclusion problem. The accuracy of the approximation controls the sparsity. We show that the proposed approach is valid for a broad class of activation functions (ReLU, sigmoid, softmax). We propose an iterative optimization algorithm to induce sparsity whose convergence is guaranteed. Because of the algorithm flexibility, the sparsity can be ensured from partial training data in a minibatch manner. To demonstrate the effectiveness of our method, we perform experiments on various networks in different applicative contexts: image classification, speech recognition, natural language processing, and time-series forecasting.
****
Unbiased Gradient Estimation in Unrolled Computation Graphs with Persistent Evolution Strategies
Paul Vicol, Luke Metz, Jascha Sohl-Dickstein
Unrolled computation graphs arise in many scenarios, including training RNNs, tuning hyperparameters through unrolled optimization, and training learned optimizers. Current approaches to optimizing parameters in such computation graphs suffer from high variance gradients, bias, slow updates, or large memory usage. We introduce a method called Persistent Evolution Strategies (PES), which divides the computation graph into a series of truncated unrolls, and performs an evolutio

n strategies-based update step after each unroll. PES eliminates bias from these truncations by accumulating correction terms over the entire sequence of unrolls. PES allows for rapid parameter updates, has low memory usage, is unbiased, and has reasonable variance characteristics. We experimentally demonstrate the advantages of PES compared to several other methods for gradient estimation on synthetic tasks, and show its applicability to training learned optimizers and tuning hyperparameters.
****

Online Graph Dictionary Learning
Cédric Vincent-Cuaz, Titouan Vayer, Rémi Flamary, Marco Corneli, Nicolas Courty
Dictionary learning is a key tool for representation learning, that explains the data as linear combination of few basic elements. Yet, this analysis is not amenable in the context of graph learning, as graphs usually belong to different metric spaces. We fill this gap by proposing a new online Graph Dictionary Learning approach, which uses the Gromov Wasserstein divergence for the data fitting term. In our work, graphs are encoded through their nodes' pairwise relations and modeled as convex combination of graph atoms, i.e. dictionary elements, estimated thanks to an online stochastic algorithm, which operates on a dataset of unregistered graphs with potentially different number of nodes. Our approach naturally extends to labeled graphs, and is completed by a novel upper bound that can be used as a fast approximation of Gromov Wasserstein in the embedding space. We provide numerical evidences showing the interest of our approach for unsupervised embedding of graph datasets and for online graph subspace estimation and tracking.
****

Neuro-algorithmic Policies Enable Fast Combinatorial Generalization
Marin Vlastelica, Michal Rolinek, Georg Martius
Although model-based and model-free approaches to learning the control of systems have achieved impressive results on standard benchmarks, generalization to task variations is still lacking. Recent results suggest that generalization for standard architectures improves only after obtaining exhaustive amounts of data. We give evidence that generalization capabilities are in many cases bottlenecked by the inability to generalize on the combinatorial aspects of the problem. We show that, for a certain subclass of the MDP framework, this can be alleviated by a neuro-algorithmic policy architecture that embeds a time-dependent shortest path solver in a deep neural network. Trained end-to-end via blackbox-differentiation, this method leads to considerable improvement in generalization capabilities in the low-data regime.
****

Efficient Training of Robust Decision Trees Against Adversarial Examples
Daniël Vos, Sicco Verwer
Current state-of-the-art algorithms for training robust decision trees have high runtime costs and require hours to run. We present GROOT, an efficient algorithm for training robust decision trees and random forests that runs in a matter of seconds to minutes. Where before the worst-case Gini impurity was computed iteratively, we find that we can solve this function analytically to improve time complexity from O(n) to O(1) in terms of n samples. Our results on both single trees and ensembles on 14 structured datasets as well as on MNIST and Fashion-MNIST demonstrate that GROOT runs several orders of magnitude faster than the state-of-the-art works and also shows better performance in terms of adversarial accuracy on structured data.
****

Object Segmentation Without Labels with Large-Scale Generative Models
Andrey Voynov, Stanislav Morozov, Artem Babenko
The recent rise of unsupervised and self-supervised learning has dramatically reduced the dependency on labeled data, providing high-quality representations for transfer on downstream tasks. Furthermore, recent works also employed these representations in a fully unsupervised setup for image classification, reducing the need for human labels on the fine-tuning stage as well. This work demonstrates that large-scale unsupervised models can also perform a more challenging object

segmentation task, requiring neither pixel-level nor image-level labeling. Namely, we show that recent unsupervised GANs allow to differentiate between foreground/background pixels, providing high-quality saliency masks. By extensive comparison on common benchmarks, we outperform existing unsupervised alternatives for object segmentation, achieving new state-of-the-art.
****
Principal Component Hierarchy for Sparse Quadratic Programs
Robbie Vreugdenhil, Viet Anh Nguyen, Armin Eftekhari, Peyman Mohajerin Esfahani
We propose a novel approximation hierarchy for cardinality-constrained, convex quadratic programs that exploits the rank-dominating eigenvectors of the quadratic matrix. Each level of approximation admits a min-max characterization whose objective function can be optimized over the binary variables analytically, while preserving convexity in the continuous variables. Exploiting this property, we propose two scalable optimization algorithms, coined as the “best response" and the “dual program", that can efficiently screen the potential indices of the nonzero elements of the original program. We show that the proposed methods are competitive with the existing screening methods in the current sparse regression literature, and it is particularly fast on instances with high number of measurements in experiments with both synthetic and real datasets.
****
Whitening and Second Order Optimization Both Make Information in the Dataset Unusable During Training, and Can Reduce or Prevent Generalization
Neha Wadia, Daniel Duckworth, Samuel S Schoenholz, Ethan Dyer, Jascha Sohl-Dickstein
Machine learning is predicated on the concept of generalization: a model achieving low error on a sufficiently large training set should also perform well on novel samples from the same distribution. We show that both data whitening and second order optimization can harm or entirely prevent generalization. In general, model training harnesses information contained in the sample-sample second moment matrix of a dataset. For a general class of models, namely models with a fully connected first layer, we prove that the information contained in this matrix is the only information which can be used to generalize. Models trained using whitened data, or with certain second order optimization schemes, have less access to this information, resulting in reduced or nonexistent generalization ability. We experimentally verify these predictions for several architectures, and further demonstrate that generalization continues to be harmed even when theoretical requirements are relaxed. However, we also show experimentally that regularized second order optimization can provide a practical tradeoff, where training is accelerated but less information is lost, and generalization can in some circumstances even improve.
****
Safe Reinforcement Learning Using Advantage-Based Intervention
Nolan C Wagener, Byron Boots, Ching-An Cheng
Many sequential decision problems involve finding a policy that maximizes total reward while obeying safety constraints. Although much recent research has focused on the development of safe reinforcement learning (RL) algorithms that produce a safe policy after training, ensuring safety during training as well remains an open problem. A fundamental challenge is performing exploration while still satisfying constraints in an unknown Markov decision process (MDP). In this work, we address this problem for the chance-constrained setting.We propose a new algorithm, SAILR, that uses an intervention mechanism based on advantage functions to keep the agent safe throughout training and optimizes the agent’s policy using off-the-shelf RL algorithms designed for unconstrained MDPs. Our method comes with strong guarantees on safety during "both" training and deployment (i.e., after training and without the intervention mechanism) and policy performance compared to the optimal safety-constrained policy. In our experiments, we show that SAILR violates constraints far less during training than standard safe RL and constrained MDP approaches and converges to a well-performing policy that can be deployed safely without intervention. Our code is available at https://github.com/nolanwagener/safe_rl.

****
Task-Optimal Exploration in Linear Dynamical Systems
Andrew J Wagenmaker, Max Simchowitz, Kevin Jamieson
Exploration in unknown environments is a fundamental problem in reinforcement learning and control. In this work, we study task-guided exploration and determine what precisely an agent must learn about their environment in order to complete a particular task. Formally, we study a broad class of decision-making problems in the setting of linear dynamical systems, a class that includes the linear quadratic regulator problem. We provide instance- and task-dependent lower bounds which explicitly quantify the difficulty of completing a task of interest. Motivated by our lower bound, we propose a computationally efficient experiment-design based exploration algorithm. We show that it optimally explores the environment, collecting precisely the information needed to complete the task, and provide finite-time bounds guaranteeing that it achieves the instance- and task-optimal sample complexity, up to constant factors. Through several examples of the linear quadratic regulator problem, we show that performing task-guided exploration provably improves on exploration schemes which do not take into account the task of interest. Along the way, we establish that certainty equivalence decision making is instance- and task-optimal, and obtain the first algorithm for the linear quadratic regulator problem which is instance-optimal. We conclude with several experiments illustrating the effectiveness of our approach in practice.
****
Learning and Planning in Average-Reward Markov Decision Processes
Yi Wan, Abhishek Naik, Richard S Sutton
We introduce learning and planning algorithms for average-reward MDPs, including 1) the first general proven-convergent off-policy model-free control algorithm without reference states, 2) the first proven-convergent off-policy model-free prediction algorithm, and 3) the first off-policy learning algorithm that converges to the actual value function rather than to the value function plus an offset. All of our algorithms are based on using the temporal-difference error rather than the conventional error when updating the estimate of the average reward. Our proof techniques are a slight generalization of those by Abounadi, Bertsekas, and Borkar (2001). In experiments with an Access-Control Queuing Task, we show some of the difficulties that can arise when using methods that rely on reference states and argue that our new algorithms are significantly easier to use.
****
Think Global and Act Local: Bayesian Optimisation over High-Dimensional Categorical and Mixed Search Spaces
Xingchen Wan, Vu Nguyen, Huong Ha, Binxin Ru, Cong Lu, Michael A. Osborne
High-dimensional black-box optimisation remains an important yet notoriously challenging problem. Despite the success of Bayesian optimisation methods on continuous domains, domains that are categorical, or that mix continuous and categorical variables, remain challenging. We propose a novel solution—we combine local optimisation with a tailored kernel design, effectively handling high-dimensional categorical and mixed search spaces, whilst retaining sample efficiency. We further derive convergence guarantee for the proposed approach. Finally, we demonstrate empirically that our method outperforms the current baselines on a variety of synthetic and real-world tasks in terms of performance, computational costs, or both.
****
Zero-Shot Knowledge Distillation from a Decision-Based Black-Box Model
Zi Wang
Knowledge distillation (KD) is a successful approach for deep neural network acceleration, with which a compact network (student) is trained by mimicking the softmax output of a pre-trained high-capacity network (teacher). In tradition, KD usually relies on access to the training samples and the parameters of the white-box teacher to acquire the transferred knowledge. However, these prerequisites are not always realistic due to storage costs or privacy issues in real-world applications. Here we propose the concept of decision-based black-box (DB3) knowledge distillation, with which the student is trained by distilling the knowledge

from a black-box teacher (parameters are not accessible) that only returns classes rather than softmax outputs. We start with the scenario when the training set is accessible. We represent a sample’s robustness against other classes by computing its distances to the teacher’s decision boundaries and use it to construct the soft label for each training sample. After that, the student can be trained via standard KD. We then extend this approach to a more challenging scenario in which even accessing the training data is not feasible. We propose to generate pseudo samples that are distinguished by the decision boundaries of the DB3 teacher to the largest extent and construct soft labels for these samples, which are used as the transfer set. We evaluate our approaches on various benchmark networks and datasets and experiment results demonstrate their effectiveness.
****

Fairness of Exposure in Stochastic Bandits

Lequn Wang, Yiwei Bai, Wen Sun, Thorsten Joachims

Contextual bandit algorithms have become widely used for recommendation in online systems (e.g. marketplaces, music streaming, news), where they now wield substantial influence on which items get shown to users. This raises questions of fairness to the items — and to the sellers, artists, and writers that benefit from this exposure. We argue that the conventional bandit formulation can lead to an undesirable and unfair winner-takes-all allocation of exposure. To remedy this problem, we propose a new bandit objective that guarantees merit-based fairness of exposure to the items while optimizing utility to the users. We formulate fairness regret and reward regret in this setting and present algorithms for both stochastic multi-armed bandits and stochastic linear bandits. We prove that the algorithms achieve sublinear fairness regret and reward regret. Beyond the theoretical analysis, we also provide empirical evidence that these algorithms can allocate exposure to different arms effectively.
****

A Proxy Variable View of Shared Confounding

Yixin Wang, David Blei

Causal inference from observational data can be biased by unobserved confounders. Confounders{—}the variables that affect both the treatments and the outcome{—}induce spurious non-causal correlations between the two. Without additional conditions, unobserved confounders generally make causal quantities hard to identify. In this paper, we focus on the setting where there are many treatments with shared confounding, and we study under what conditions is causal identification possible. The key observation is that we can view subsets of treatments as proxies of the unobserved confounder and identify the intervention distributions of the rest. Moreover, while existing identification formulas for proxy variables involve solving integral equations, we show that one can circumvent the need for such solutions by directly modeling the data. Finally, we extend these results to an expanded class of causal graphs, those with other confounders and selection variables.
****

Fast Algorithms for Stackelberg Prediction Game with Least Squares Loss

Jiali Wang, He Chen, Rujun Jiang, Xudong Li, Zihao Li

The Stackelberg prediction game (SPG) has been extensively used to model the interactions between the learner and data provider in the training process of various machine learning algorithms. Particularly, SPGs played prominent roles in cybersecurity applications, such as intrusion detection, banking fraud detection, spam filtering, and malware detection. Often formulated as NP-hard bi-level optimization problems, it is generally computationally intractable to find global solutions to SPGs. As an interesting progress in this area, a special class of SPGs with the least squares loss (SPG-LS) have recently been shown polynomially solvable by a bisection method. However, in each iteration of this method, a semidefinite program (SDP) needs to be solved. The resulted high computational costs prevent its applications for large-scale problems. In contrast, we propose a novel approach that reformulates a SPG-LS as a single SDP of a similar form and the same dimension as those solved in the bisection method. Our SDP reformulation is, evidenced by our numerical experiments, orders of magnitude faster than the exi

sting bisection method. We further show that the obtained SDP can be reduced to a second order cone program (SOCP). This allows us to provide real-time response to large-scale SPG-LS problems. Numerical results on both synthetic and real world datasets indicate that the proposed SOCP method is up to 20,000+ times faster than the state of the art.
****
Accelerate CNNs from Three Dimensions: A Comprehensive Pruning Framework
Wenxiao Wang, Minghao Chen, Shuai Zhao, Long Chen, Jinming Hu, Haifeng Liu, Deng Cai, Xiaofei He, Wei Liu
Most neural network pruning methods, such as filter-level and layer-level prunings, prune the network model along one dimension (depth, width, or resolution) solely to meet a computational budget. However, such a pruning policy often leads to excessive reduction of that dimension, thus inducing a huge accuracy loss. To alleviate this issue, we argue that pruning should be conducted along three dimensions comprehensively. For this purpose, our pruning framework formulates pruning as an optimization problem. Specifically, it first casts the relationships between a certain model’s accuracy and depth/width/resolution into a polynomial regression and then maximizes the polynomial to acquire the optimal values for the three dimensions. Finally, the model is pruned along the three optimal dimensions accordingly. In this framework, since collecting too much data for training the regression is very time-costly, we propose two approaches to lower the cost: 1) specializing the polynomial to ensure an accurate regression even with less training data; 2) employing iterative pruning and fine-tuning to collect the data faster. Extensive experiments show that our proposed algorithm surpasses state-of-the-art pruning algorithms and even neural architecture search-based algorithms.
****
Explainable Automated Graph Representation Learning with Hyperparameter Importance
Xin Wang, Shuyi Fan, Kun Kuang, Wenwu Zhu
Current graph representation (GR) algorithms require huge demand of human experts in hyperparameter tuning, which significantly limits their practical applications, leading to an urge for automated graph representation without human intervention. Although automated machine learning (AutoML) serves as a good candidate for automatic hyperparameter tuning, little literature has been reported on automated graph presentation learning and the only existing work employs a black-box strategy, lacking insights into explaining the relative importance of different hyperparameters. To address this issue, we study explainable automated graph representation with hyperparameter importance in this paper. We propose an explainable AutoML approach for graph representation (e-AutoGR) which utilizes explainable graph features during performance estimation and learns decorrelated importance weights for different hyperparameters in affecting the model performance through a non-linear decorrelated weighting regression. These learned importance weights can in turn help to provide more insights in hyperparameter search procedure. We theoretically prove the soundness of the decorrelated weighting algorithm. Extensive experiments on real-world datasets demonstrate the superiority of our proposed e-AutoGR model against state-of-the-art methods in terms of both model performance and hyperparameter importance explainability.
****
Self-Tuning for Data-Efficient Deep Learning
Ximei Wang, Jinghan Gao, Mingsheng Long, Jianmin Wang
Deep learning has made revolutionary advances to diverse applications in the presence of large-scale labeled datasets. However, it is prohibitively time-costly and labor-expensive to collect sufficient labeled data in most realistic scenarios. To mitigate the requirement for labeled data, semi-supervised learning (SSL) focuses on simultaneously exploring both labeled and unlabeled data, while transfer learning (TL) popularizes a favorable practice of fine-tuning a pre-trained model to the target data. A dilemma is thus encountered: Without a decent pre-trained model to provide an implicit regularization, SSL through self-training from scratch will be easily misled by inaccurate pseudo-labels, especially in larg

e-sized label space; Without exploring the intrinsic structure of unlabeled data , TL through fine-tuning from limited labeled data is at risk of under-transfer caused by model shift. To escape from this dilemma, we present Self-Tuning to enable data-efficient deep learning by unifying the exploration of labeled and unlabeled data and the transfer of a pre-trained model, as well as a Pseudo Group Contrast (PGC) mechanism to mitigate the reliance on pseudo-labels and boost the tolerance to false labels. Self-Tuning outperforms its SSL and TL counterparts on five tasks by sharp margins, e.g. it doubles the accuracy of fine-tuning on Cars with $15%$ labels.
****
Label Distribution Learning Machine
Jing Wang, Xin Geng
Although Label Distribution Learning (LDL) has witnessed extensive classification applications, it faces the challenge of objective mismatch – the objective of LDL mismatches that of classification, which has seldom been noticed in existing studies. Our goal is to solve the objective mismatch and improve the classification performance of LDL. Specifically, we extend the margin theory to LDL and propose a new LDL method called \textbf{L}abel \textbf{D}istribution \textbf{L}earning \textbf{M}achine (LDLM). First, we define the label distribution margin and propose the \textbf{S}upport \textbf{V}ector \textbf{R}egression \textbf{M}achine (SVRM) to learn the optimal label. Second, we propose the adaptive margin loss to learn label description degrees. In theoretical analysis, we develop a generalization theory for the SVRM and analyze the generalization of LDLM. Experimental results validate the better classification performance of LDLM.
****
AlphaNet: Improved Training of Supernets with Alpha-Divergence
Dilin Wang, Chengyue Gong, Meng Li, Qiang Liu, Vikas Chandra
Weight-sharing neural architecture search (NAS) is an effective technique for automating efficient neural architecture design. Weight-sharing NAS builds a supernet that assembles all the architectures as its sub-networks and jointly trains the supernet with the sub-networks. The success of weight-sharing NAS heavily relies on distilling the knowledge of the supernet to the sub-networks. However, we find that the widely used distillation divergence, i.e., KL divergence, may lead to student sub-networks that over-estimate or under-estimate the uncertainty of the teacher supernet, leading to inferior performance of the sub-networks. In this work, we propose to improve the supernet training with a more generalized alpha-divergence. By adaptively selecting the alpha-divergence, we simultaneously prevent the over-estimation or under-estimation of the uncertainty of the teacher model. We apply the proposed alpha-divergence based supernets training to both slimmable neural networks and weight-sharing NAS, and demonstrate significant improvements. Specifically, our discovered model family, AlphaNet, outperforms prior-art models on a wide range of FLOPs regimes, including BigNAS, Once-for-All networks, and AttentiveNAS. We achieve ImageNet top-1 accuracy of 80.0% with only 444M FLOPs. Our code and pretrained models are available at https://github.com/facebookresearch/AlphaNet.
****
Global Convergence of Policy Gradient for Linear-Quadratic Mean-Field Control/Game in Continuous Time
Weichen Wang, Jiequn Han, Zhuoran Yang, Zhaoran Wang
Recent years have witnessed the success of multi-agent reinforcement learning, which has motivated new research directions for mean-field control (MFC) and mean-field game (MFG), as the multi-agent system can be well approximated by a mean-field problem when the number of agents grows to be very large. In this paper, we study the policy gradient (PG) method for the linear-quadratic mean-field control and game, where we assume each agent has identical linear state transitions and quadratic cost functions. While most recent works on policy gradient for MFC and MFG are based on discrete-time models, we focus on a continuous-time model where some of our analyzing techniques could be valuable to the interested readers. For both the MFC and the MFG, we provide PG update and show that it converges to the optimal solution at a linear rate, which is verified by a synthetic sim

ulation. For the MFG, we also provide sufficient conditions for the existence and uniqueness of the Nash equilibrium.
****
SG-PALM: a Fast Physically Interpretable Tensor Graphical Model
Yu Wang, Alfred Hero
We propose a new graphical model inference procedure, called SG-PALM, for learning conditional dependency structure of high-dimensional tensor-variate data. Unlike most other tensor graphical models the proposed model is interpretable and computationally scalable to high dimension. Physical interpretability follows from the Sylvester generative (SG) model on which SG-PALM is based: the model is exact for any observation process that is a solution of a partial differential equation of Poisson type. Scalability follows from the fast proximal alternating linearized minimization (PALM) procedure that SG-PALM uses during training. We establish that SG-PALM converges linearly (i.e., geometric convergence rate) to a global optimum of its objective function. We demonstrate scalability and accuracy of SG-PALM for an important but challenging climate prediction problem: spatio-temporal forecasting of solar flares from multimodal imaging data.
****
Deep Generative Learning via Schrödinger Bridge
Gefei Wang, Yuling Jiao, Qian Xu, Yang Wang, Can Yang
We propose to learn a generative model via entropy interpolation with a Schr{ö}dinger Bridge. The generative learning task can be formulated as interpolating between a reference distribution and a target distribution based on the Kullback-Leibler divergence. At the population level, this entropy interpolation is characterized via an SDE on [0,1] with a time-varying drift term. At the sample level, we derive our Schr{ö}dinger Bridge algorithm by plugging the drift term estimated by a deep score estimator and a deep density ratio estimator into the Euler-Maruyama method. Under some mild smoothness assumptions of the target distribution, we prove the consistency of both the score estimator and the density ratio estimator, and then establish the consistency of the proposed Schr{ö}dinger Bridge approach. Our theoretical results guarantee that the distribution learned by our approach converges to the target distribution. Experimental results on multimodal synthetic data and benchmark data support our theoretical findings and indicate that the generative model via Schr{ö}dinger Bridge is comparable with state-of-the-art GANs, suggesting a new formulation of generative learning. We demonstrate its usefulness in image interpolation and image inpainting.
****
Robust Inference for High-Dimensional Linear Models via Residual Randomization
Y. Samuel Wang, Si Kai Lee, Panos Toulis, Mladen Kolar
We propose a residual randomization procedure designed for robust inference using Lasso estimates in the high-dimensional setting. Compared to earlier work that focuses on sub-Gaussian errors, the proposed procedure is designed to work robustly in settings that also include heavy-tailed covariates and errors. Moreover, our procedure can be valid under clustered errors, which is important in practice, but has been largely overlooked by earlier work. Through extensive simulations, we illustrate our method’s wider range of applicability as suggested by theory. In particular, we show that our method outperforms state-of-art methods in challenging, yet more realistic, settings where the distribution of covariates is heavy-tailed or the sample size is small, while it remains competitive in standard, “well behaved" settings previously studied in the literature.
****
A Modular Analysis of Provable Acceleration via Polyak’s Momentum: Training a Wide ReLU Network and a Deep Linear Network
Jun-Kun Wang, Chi-Heng Lin, Jacob D Abernethy
Incorporating a so-called “momentum” dynamic in gradient descent methods is widely used in neural net training as it has been broadly observed that, at least empirically, it often leads to significantly faster convergence. At the same time, there are very few theoretical guarantees in the literature to explain this apparent acceleration effect. Even for the classical strongly convex quadratic problems, several existing results only show Polyak’s momentum has an accelerated li

near rate asymptotically. In this paper, we first revisit the quadratic problems and show a non-asymptotic accelerated linear rate of Polyak’s momentum. Then, we provably show that Polyak’s momentum achieves acceleration for training a one-layer wide ReLU network and a deep linear network, which are perhaps the two most popular canonical models for studying optimization and deep learning in the literature. Prior works (Du et al. 2019) and (Wu et al. 2019) showed that using vanilla gradient descent, and with an use of over-parameterization, the error decays as $(1- \Theta(\frac{1}{ \kappa'}))^t$ after t iterations, where κ' is the condition number of a Gram Matrix. Our result shows that with the appropriate choice of parameters Polyak’s momentum has a rate of $(1-\Theta(\frac{1}{\sqrt{\kappa'}}))^t$. For the deep linear network, prior work (Hu et al. 2020) showed that vanilla gradient descent has a rate of $(1-\Theta(\frac{1}{\kappa}))^t$, where κ is the condition number of a data matrix. Our result shows an acceleration rate $(1- \Theta(\frac{1}{\sqrt{\kappa}}))^t$ is achievable by Polyak’s momentum. This work establishes that momentum does indeed speed up neural net training.
****
Optimal Non-Convex Exact Recovery in Stochastic Block Model via Projected Power Method
Peng Wang, Huikang Liu, Zirui Zhou, Anthony Man-Cho So
In this paper, we study the problem of exact community recovery in the symmetric stochastic block model, where a graph of n vertices is randomly generated by partitioning the vertices into $K \ge 2$ equal-sized communities and then connecting each pair of vertices with probability that depends on their community memberships. Although the maximum-likelihood formulation of this problem is discrete and non-convex, we propose to tackle it directly using projected power iterations with an initialization that satisfies a partial recovery condition. Such an initialization can be obtained by a host of existing methods. We show that in the logarithmic degree regime of the considered problem, the proposed method can exactly recover the underlying communities at the information-theoretic limit. Moreover, with a qualified initialization, it runs in $\mO(n\log^2n/\log\log n)$ time, which is competitive with existing state-of-the-art methods. We also present numerical results of the proposed method to support and complement our theoretical development.
****
ConvexVST: A Convex Optimization Approach to Variance-stabilizing Transformation
Mengfan Wang, Boyu Lyu, Guoqiang Yu
The variance-stabilizing transformation (VST) problem is to transform heteroscedastic data to homoscedastic data so that they are more tractable for subsequent analysis. However, most of the existing approaches focus on finding an analytical solution for a certain parametric distribution, which severely limits the applications, because simple distributions cannot faithfully describe the real data while more complicated distributions cannot be analytically solved. In this paper, we converted the VST problem into a convex optimization problem, which can always be efficiently solved, identified the specific structure of the convex problem, which further improved the efficiency of the proposed algorithm, and showed that any finite discrete distributions and the discretized version of any continuous distributions from real data can be variance-stabilized in an easy and nonparametric way. We demonstrated the new approach on bioimaging data and achieved superior performance compared to peer algorithms in terms of not only the variance homoscedasticity but also the impact on subsequent analysis such as denoising. Source codes are available at https://github.com/yu-lab-vt/ConvexVST.
****
The Implicit Bias for Adaptive Optimization Algorithms on Homogeneous Neural Networks
Bohan Wang, Qi Meng, Wei Chen, Tie-Yan Liu
Despite their overwhelming capacity to overfit, deep neural networks trained by specific optimization algorithms tend to generalize relatively well to unseen data. Recently, researchers explained it by investigating the implicit bias of optimization algorithms. A remarkable progress is the work (Lyu & Li, 2019), which

proves gradient descent (GD) maximizes the margin of homogeneous deep neural networks. Except the first-order optimization algorithms like GD, adaptive algorithms such as AdaGrad, RMSProp and Adam are popular owing to their rapid training process. Mean-while, numerous works have provided empirical evidence that adaptive methods may suffer from poor generalization performance. However, theoretical explanation for the generalization of adaptive optimization algorithms is still lacking. In this paper, we study the implicit bias of adaptive optimization algorithms on homogeneous neural networks. In particular, we study the convergent direction of parameters when they are optimizing the logistic loss. We prove that the convergent direction of Adam and RMSProp is the same as GD, while for AdaGrad, the convergent direction depends on the adaptive conditioner. Technically, we provide a unified framework to analyze convergent direction of adaptive optimization algorithms by constructing novel and nontrivial adaptive gradient flow and surrogate margin. The theoretical findings explain the superiority on generalization of exponential moving average strategy that is adopted by RMSProp and Adam. To the best of knowledge, it is the first work to study the convergent direction of adaptive optimizations on non-linear deep neural networks
****
Robust Learning for Data Poisoning Attacks
Yunjuan Wang, Poorya Mianjy, Raman Arora
We investigate the robustness of stochastic approximation approaches against data poisoning attacks. We focus on two-layer neural networks with ReLU activation and show that under a specific notion of separability in the RKHS induced by the infinite-width network, training (finite-width) networks with stochastic gradient descent is robust against data poisoning attacks. Interestingly, we find that in addition to a lower bound on the width of the network, which is standard in the literature, we also require a distribution-dependent upper bound on the width for robust generalization. We provide extensive empirical evaluations that support and validate our theoretical results.
****
SketchEmbedNet: Learning Novel Concepts by Imitating Drawings
Alexander Wang, Mengye Ren, Richard Zemel
Sketch drawings capture the salient information of visual concepts. Previous work has shown that neural networks are capable of producing sketches of natural objects drawn from a small number of classes. While earlier approaches focus on generation quality or retrieval, we explore properties of image representations learned by training a model to produce sketches of images. We show that this generative, class-agnostic model produces informative embeddings of images from novel examples, classes, and even novel datasets in a few-shot setting. Additionally, we find that these learned representations exhibit interesting structure and compositionality.
****
Directional Bias Amplification
Angelina Wang, Olga Russakovsky
Mitigating bias in machine learning systems requires refining our understanding of bias propagation pathways: from societal structures to large-scale data to trained models to impact on society. In this work, we focus on one aspect of the problem, namely bias amplification: the tendency of models to amplify the biases present in the data they are trained on. A metric for measuring bias amplification was introduced in the seminal work by Zhao et al. (2017); however, as we demonstrate, this metric suffers from a number of shortcomings including conflating different types of bias amplification and failing to account for varying base rates of protected attributes. We introduce and analyze a new, decoupled metric for measuring bias amplification, $BiasAmp_{\rightarrow}$ (Directional Bias Amplification). We thoroughly analyze and discuss both the technical assumptions and normative implications of this metric. We provide suggestions about its measurement by cautioning against predicting sensitive attributes, encouraging the use of confidence intervals due to fluctuations in the fairness of models across runs, and discussing the limitations of what this metric captures. Throughout this paper, we work to provide an interrogative look at the technical measurement of bi

as amplification, guided by our normative ideas of what we want it to encompass. Code is located at https://github.com/princetonvisualai/directional-bias-amp.
****
An exact solver for the Weston-Watkins SVM subproblem
Yutong Wang, Clayton Scott
Recent empirical evidence suggests that the Weston-Watkins support vector machine is among the best performing multiclass extensions of the binary SVM. Current state-of-the-art solvers repeatedly solve a particular subproblem approximately using an iterative strategy. In this work, we propose an algorithm that solves the subproblem exactly using a novel reparametrization of the Weston-Watkins dual problem. For linear WW-SVMs, our solver shows significant speed-up over the state-of-the-art solver when the number of classes is large. Our exact subproblem solver also allows us to prove linear convergence of the overall solver.
****
SCC: an efficient deep reinforcement learning agent mastering the game of StarCraft II
Xiangjun Wang, Junxiao Song, Penghui Qi, Peng Peng, Zhenkun Tang, Wei Zhang, Weimin Li, Xiongjun Pi, Jujie He, Chao Gao, Haitao Long, Quan Yuan
AlphaStar, the AI that reaches GrandMaster level in StarCraft II, is a remarkable milestone demonstrating what deep reinforcement learning can achieve in complex Real-Time Strategy (RTS) games. However, the complexities of the game, algorithms and systems, and especially the tremendous amount of computation needed are big obstacles for the community to conduct further research in this direction. We propose a deep reinforcement learning agent, StarCraft Commander (SCC). With order of magnitude less computation, it demonstrates top human performance defeating GrandMaster players in test matches and top professional players in a live event. Moreover, it shows strong robustness to various human strategies and discovers novel strategies unseen from human plays. In this paper, we’ll share the key insights and optimizations on efficient imitation learning and reinforcement learning for StarCraft II full game.
****
Quantum algorithms for reinforcement learning with a generative model
Daochen Wang, Aarthi Sundaram, Robin Kothari, Ashish Kapoor, Martin Roetteler
Reinforcement learning studies how an agent should interact with an environment to maximize its cumulative reward. A standard way to study this question abstractly is to ask how many samples an agent needs from the environment to learn an optimal policy for a γ-discounted Markov decision process (MDP). For such an MDP, we design quantum algorithms that approximate an optimal policy (π^*), the optimal value function (v^*), and the optimal Q-function (q^*), assuming the algorithms can access samples from the environment in quantum superposition. This assumption is justified whenever there exists a simulator for the environment; for example, if the environment is a video game or some other program. Our quantum algorithms, inspired by value iteration, achieve quadratic speedups over the best-possible classical sample complexities in the approximation accuracy (ϵ) and two main parameters of the MDP: the effective time horizon ($\frac{1}{1-\gamma}$) and the size of the action space (A). Moreover, we show that our quantum algorithm for computing q^* is optimal by proving a matching quantum lower bound.
****
Matrix Completion with Model-free Weighting
Jiayi Wang, Raymond K. W. Wong, Xiaojun Mao, Kwun Chuen Gary Chan
In this paper, we propose a novel method for matrix completion under general non-uniform missing structures. By controlling an upper bound of a novel balancing error, we construct weights that can actively adjust for the non-uniformity in the empirical risk without explicitly modeling the observation probabilities, and can be computed efficiently via convex optimization. The recovered matrix based on the proposed weighted empirical risk enjoys appealing theoretical guarantees. In particular, the proposed method achieves stronger guarantee than existing work in terms of the scaling with respect to the observation probabilities, under asymptotically heterogeneous missing settings (where entry-wise observation pro

babilities can be of different orders). These settings can be regarded as a bett er theoretical model of missing patterns with highly varying probabilities. We a lso provide a new minimax lower bound under a class of heterogeneous settings. N umerical experiments are also provided to demonstrate the effectiveness of the p roposed method.
****
UniSpeech: Unified Speech Representation Learning with Labeled and Unlabeled Dat a
Chengyi Wang, Yu Wu, Yao Qian, Kenichi Kumatani, Shujie Liu, Furu Wei, Michael Z eng, Xuedong Huang
In this paper, we propose a unified pre-training approach called UniSpeech to le arn speech representations with both labeled and unlabeled data, in which superv ised phonetic CTC learning and phonetically-aware contrastive self-supervised le arning are conducted in a multi-task learning manner. The resultant representati ons can capture information more correlated with phonetic structures and improve the generalization across languages and domains. We evaluate the effectiveness of UniSpeech for cross-lingual representation learning on public CommonVoice cor pus. The results show that UniSpeech outperforms self-supervised pretraining and supervised transfer learning for speech recognition by a maximum of 13.4% and 2 6.9% relative phone error rate reductions respectively (averaged over all testin g languages). The transferability of UniSpeech is also verified on a domain-shif t speech recognition task, i.e., a relative word error rate reduction of 6% agai nst the previous approach.
****
Instabilities of Offline RL with Pre-Trained Neural Representation
Ruosong Wang, Yifan Wu, Ruslan Salakhutdinov, Sham Kakade
In offline reinforcement learning (RL), we seek to utilize offline data to evalu ate (or learn) policies in scenarios where the data are collected from a distrib ution that substantially differs from that of the target policy to be evaluated. Recent theoretical advances have shown that such sample-efficient offline RL is indeed possible provided certain strong representational conditions hold, else there are lower bounds exhibiting exponential error amplification (in the proble m horizon) unless the data collection distribution has only a mild distribution shift relative to the target policy. This work studies these issues from an empi rical perspective to gauge how stable offline RL methods are. In particular, our methodology explores these ideas when using features from pre-trained neural ne tworks, in the hope that these representations are powerful enough to permit sam ple efficient offline RL. Through extensive experiments on a range of tasks, we see that substantial error amplification does occur even when using such pre-tra ined representations (trained on the same task itself); we find offline RL is st able only under extremely mild distribution shift. The implications of these res ults, both from a theoretical and an empirical perspective, are that successful offline RL (where we seek to go beyond the low distribution shift regime) requir es substantially stronger conditions beyond those which suffice for successful s upervised learning.
****
Learning to Weight Imperfect Demonstrations
Yunke Wang, Chang Xu, Bo Du, Honglak Lee
This paper investigates how to weight imperfect expert demonstrations for genera tive adversarial imitation learning (GAIL). The agent is expected to perform beh aviors demonstrated by experts. But in many applications, experts could also mak e mistakes and their demonstrations would mislead or slow the learning process o f the agent. Recently, existing methods for imitation learning from imperfect de monstrations mostly focus on using the preference or confidence scores to distin guish imperfect demonstrations. However, these auxiliary information needs to be collected with the help of an oracle, which is usually hard and expensive to af ford in practice. In contrast, this paper proposes a method of learning to weigh t imperfect demonstrations in GAIL without imposing extensive prior information. We provide a rigorous mathematical analysis, presenting that the weights of dem onstrations can be exactly determined by combining the discriminator and agent p

olicy in GAIL. Theoretical analysis suggests that with the estimated weights the agent can learn a better policy beyond those plain expert demonstrations. Experiments in the Mujoco and Atari environments demonstrate that the proposed algorithm outperforms baseline methods in handling imperfect expert demonstrations.
****

Evolving Attention with Residual Convolutions

Yujing Wang, Yaming Yang, Jiangang Bai, Mingliang Zhang, Jing Bai, Jing Yu, Ce Zhang, Gao Huang, Yunhai Tong

Transformer is a ubiquitous model for natural language processing and has attracted wide attentions in computer vision. The attention maps are indispensable for a transformer model to encode the dependencies among input tokens. However, they are learned independently in each layer and sometimes fail to capture precise patterns. In this paper, we propose a novel and generic mechanism based on evolving attention to improve the performance of transformers. On one hand, the attention maps in different layers share common knowledge, thus the ones in preceding layers can instruct the attention in succeeding layers through residual connections. On the other hand, low-level and high-level attentions vary in the level of abstraction, so we adopt convolutional layers to model the evolutionary process of attention maps. The proposed evolving attention mechanism achieves significant performance improvement over various state-of-the-art models for multiple tasks, including image classification, natural language understanding and machine translation.
****

Guarantees for Tuning the Step Size using a Learning-to-Learn Approach

Xiang Wang, Shuai Yuan, Chenwei Wu, Rong Ge

Choosing the right parameters for optimization algorithms is often the key to their success in practice. Solving this problem using a learning-to-learn approach—using meta-gradient descent on a meta-objective based on the trajectory that the optimizer generates—was recently shown to be effective. However, the meta-optimization problem is difficult. In particular, the meta-gradient can often explode/vanish, and the learned optimizer may not have good generalization performance if the meta-objective is not chosen carefully. In this paper we give meta-optimization guarantees for the learning-to-learn approach on a simple problem of tuning the step size for quadratic loss. Our results show that the naïve objective suffers from meta-gradient explosion/vanishing problem. Although there is a way to design the meta-objective so that the meta-gradient remains polynomially bounded, computing the meta-gradient directly using backpropagation leads to numerical issues. We also characterize when it is necessary to compute the meta-objective on a separate validation set to ensure the generalization performance of the learned optimizer. Finally, we verify our results empirically and show that a similar phenomenon appears even for more complicated learned optimizers parametrized by neural networks.
****

Bridging Multi-Task Learning and Meta-Learning: Towards Efficient Training and Effective Adaptation

Haoxiang Wang, Han Zhao, Bo Li

Multi-task learning (MTL) aims to improve the generalization of several related tasks by learning them jointly. As a comparison, in addition to the joint training scheme, modern meta-learning allows unseen tasks with limited labels during the test phase, in the hope of fast adaptation over them. Despite the subtle difference between MTL and meta-learning in the problem formulation, both learning paradigms share the same insight that the shared structure between existing training tasks could lead to better generalization and adaptation. In this paper, we take one important step further to understand the close connection between these two learning paradigms, through both theoretical analysis and empirical investigation. Theoretically, we first demonstrate that MTL shares the same optimization formulation with a class of gradient-based meta-learning (GBML) algorithms. We then prove that for over-parameterized neural networks with sufficient depth, the learned predictive functions of MTL and GBML are close. In particular, this result implies that the predictions given by these two models are similar over th

e same unseen task. Empirically, we corroborate our theoretical findings by show ing that, with proper implementation, MTL is competitive against state-of-the-ar t GBML algorithms on a set of few-shot image classification benchmarks. Since ex isting GBML algorithms often involve costly second-order bi-level optimization, our first-order MTL method is an order of magnitude faster on large-scale datase ts such as mini-ImageNet. We believe this work could help bridge the gap between these two learning paradigms, and provide a computationally efficient alternati ve to GBML that also supports fast task adaptation.
****
Towards Better Laplacian Representation in Reinforcement Learning with Generaliz ed Graph Drawing
Kaixin Wang, Kuangqi Zhou, Qixin Zhang, Jie Shao, Bryan Hooi, Jiashi Feng
The Laplacian representation recently gains increasing attention for reinforceme nt learning as it provides succinct and informative representation for states, b y taking the eigenvectors of the Laplacian matrix of the state-transition graph as state embeddings. Such representation captures the geometry of the underlying state space and is beneficial to RL tasks such as option discovery and reward s haping. To approximate the Laplacian representation in large (or even continuous) state spaces, recent works propose to minimize a spectral graph drawing object ive, which however has infinitely many global minimizers other than the eigenvec tors. As a result, their learned Laplacian representation may differ from the gr ound truth. To solve this problem, we reformulate the graph drawing objective in to a generalized form and derive a new learning objective, which is proved to ha ve eigenvectors as its unique global minimizer. It enables learning high-quality Laplacian representations that faithfully approximate the ground truth. We vali date this via comprehensive experiments on a set of gridworld and continuous con trol environments. Moreover, we show that our learned Laplacian representations lead to more exploratory options and better reward shaping.
****
Robust Asymmetric Learning in POMDPs
Andrew Warrington, Jonathan W Lavington, Adam Scibior, Mark Schmidt, Frank Wood
Policies for partially observed Markov decision processes can be efficiently lea rned by imitating expert policies generated using asymmetric information. Unfort unately, existing approaches for this kind of imitation learning have a serious flaw: the expert does not know what the trainee cannot see, and as a result may encourage actions that are sub-optimal or unsafe under partial information. To a ddress this issue, we derive an update which, when applied iteratively to an exp ert, maximizes the expected reward of the trainee’s policy. Using this update, w e construct a computationally efficient algorithm, adaptive asymmetric DAgger (A 2D), that jointly trains the expert and trainee policies. We then show that A2D allows the trainee to safely imitate the modified expert, and outperforms polici es learned either by imitating a fixed expert or through direct reinforcement le arning.
****
A Unified Generative Adversarial Network Training via Self-Labeling and Self-Att ention
Tomoki Watanabe, Paolo Favaro
We propose a novel GAN training scheme that can handle any level of labeling in a unified manner. Our scheme introduces a form of artificial labeling that can i ncorporate manually defined labels, when available, and induce an alignment betw een them. To define the artificial labels, we exploit the assumption that neural network generators can be trained more easily to map nearby latent vectors to d ata with semantic similarities, than across separate categories. We use generate d data samples and their corresponding artificial conditioning labels to train a classifier. The classifier is then used to self-label real data. To boost the a ccuracy of the self-labeling, we also use the exponential moving average of the classifier. However, because the classifier might still make mistakes, especiall y at the beginning of the training, we also refine the labels through self-atten tion, by using the labeling of real data samples only when the classifier output s a high classification probability score. We evaluate our approach on CIFAR-10,

STL-10 and SVHN, and show that both self-labeling and self-attention consistently improve the quality of generated data. More surprisingly, we find that the proposed scheme can even outperform class-conditional GANs.
****
Decision-Making Under Selective Labels: Optimal Finite-Domain Policies and Beyond
Dennis Wei
Selective labels are a common feature of high-stakes decision-making applications, referring to the lack of observed outcomes under one of the possible decisions. This paper studies the learning of decision policies in the face of selective labels, in an online setting that balances learning costs against future utility. In the homogeneous case in which individuals' features are disregarded, the optimal decision policy is shown to be a threshold policy. The threshold becomes more stringent as more labels are collected; the rate at which this occurs is characterized. In the case of features drawn from a finite domain, the optimal policy consists of multiple homogeneous policies in parallel. For the general infinite-domain case, the homogeneous policy is extended by using a probabilistic classifier and bootstrapping to provide its inputs. In experiments on synthetic and real data, the proposed policies achieve consistently superior utility with no parameter tuning in the finite-domain case and lower parameter sensitivity in the general case.
****
Inferring serial correlation with dynamic backgrounds
Song Wei, Yao Xie, Dobromir Rahnev
Sequential data with serial correlation and an unknown, unstructured, and dynamic background is ubiquitous in neuroscience, psychology, and econometrics. Inferring serial correlation for such data is a fundamental challenge in statistics. We propose a Total Variation (TV) constrained least square estimator coupled with hypothesis tests to infer the serial correlation in the presence of unknown and unstructured dynamic background. The TV constraint on the dynamic background encourages a piecewise constant structure, which can approximate a wide range of dynamic backgrounds. The tuning parameter is selected via the Ljung-Box test to control the bias-variance trade-off. We establish a non-asymptotic upper bound for the estimation error through variational inequalities. We also derive a lower error bound via Fano's method and show the proposed method is near-optimal. Numerical simulation and a real study in psychology demonstrate the excellent performance of our proposed method compared with the state-of-the-art.
****
Meta-learning Hyperparameter Performance Prediction with Neural Processes
Ying Wei, Peilin Zhao, Junzhou Huang
The surrogate that predicts the performance of hyperparameters has been a key component for sequential model-based hyperparameter optimization. In practical applications, a trial of a hyper-parameter configuration may be so costly that a surrogate is expected to return an optimal configuration with as few trials as possible. Observing that human experts draw on their expertise in a machine learning model by trying configurations that once performed well on other datasets, we are inspired to build a trial-efficient surrogate by transferring the meta-knowledge learned from historical trials on other datasets. We propose an end-to-end surrogate named as Transfer NeuralProcesses (TNP) that learns a comprehensive set of meta-knowledge, including the parameters of historical surrogates, historical trials, and initial configurations for other datasets. Experiments on extensive OpenML datasets and three computer vision datasets demonstrate that the proposed algorithm achieves state-of-the-art performance in at least one order of magnitude less trials.
****
A Structured Observation Distribution for Generative Biological Sequence Prediction and Forecasting
Eli N Weinstein, Debora Marks
Generative probabilistic modeling of biological sequences has widespread existing and potential application across biology and biomedicine, from evolutionary bi

ology to epidemiology to protein design. Many standard sequence analysis methods preprocess data using a multiple sequence alignment (MSA) algorithm, one of the most widely used computational methods in all of science. However, as we show in this article, training generative probabilistic models with MSA preprocessing leads to statistical pathologies in the context of sequence prediction and forecasting. To address these problems, we propose a principled drop-in alternative to MSA preprocessing in the form of a structured observation distribution (the "MuE" distribution). We prove theoretically that the MuE distribution comprehensively generalizes popular methods for inferring biological sequence alignments, and provide a precise characterization of how such biological models have differed from natural language latent alignment models. We show empirically that models that use the MuE as an observation distribution outperform comparable methods across a variety of datasets, and apply MuE models to a novel problem for generative probabilistic sequence models: forecasting pathogen evolution.
****

Thinking Like Transformers

Gail Weiss, Yoav Goldberg, Eran Yahav

What is the computational model behind a Transformer? Where recurrent neural networks have direct parallels in finite state machines, allowing clear discussion and thought around architecture variants or trained models, Transformers have no such familiar parallel. In this paper we aim to change that, proposing a computational model for the transformer-encoder in the form of a programming language. We map the basic components of a transformer-encoder—attention and feed-forward computation—into simple primitives, around which we form a programming language: the Restricted Access Sequence Processing Language (RASP). We show how RASP can be used to program solutions to tasks that could conceivably be learned by a Transformer, and how a Transformer can be trained to mimic a RASP solution. In particular, we provide RASP programs for histograms, sorting, and Dyck-languages. We further use our model to relate their difficulty in terms of the number of required layers and attention heads: analyzing a RASP program implies a maximum number of heads and layers necessary to encode a task in a transformer. Finally, we see how insights gained from our abstraction might be used to explain phenomena seen in recent works.
****

Leveraged Weighted Loss for Partial Label Learning

Hongwei Wen, Jingyi Cui, Hanyuan Hang, Jiabin Liu, Yisen Wang, Zhouchen Lin

As an important branch of weakly supervised learning, partial label learning deals with data where each instance is assigned with a set of candidate labels, whereas only one of them is true. Despite many methodology studies on learning from partial labels, there still lacks theoretical understandings of their risk consistent properties under relatively weak assumptions, especially on the link between theoretical results and the empirical choice of parameters. In this paper, we propose a family of loss functions named \textit{Leveraged Weighted} (LW) loss, which for the first time introduces the leverage parameter β to consider the trade-off between losses on partial labels and non-partial ones. From the theoretical side, we derive a generalized result of risk consistency for the LW loss in learning from partial labels, based on which we provide guidance to the choice of the leverage parameter β. In experiments, we verify the theoretical guidance, and show the high effectiveness of our proposed LW loss on both benchmark and real datasets compared with other state-of-the-art partial label learning algorithms.
****

Characterizing the Gap Between Actor-Critic and Policy Gradient

Junfeng Wen, Saurabh Kumar, Ramki Gummadi, Dale Schuurmans

Actor-critic (AC) methods are ubiquitous in reinforcement learning. Although it is understood that AC methods are closely related to policy gradient (PG), their precise connection has not been fully characterized previously. In this paper, we explain the gap between AC and PG methods by identifying the exact adjustment to the AC objective/gradient that recovers the true policy gradient of the cumulative reward objective (PG). Furthermore, by viewing the AC method as a two-pla

yer Stackelberg game between the actor and critic, we show that the Stackelberg policy gradient can be recovered as a special case of our more general analysis. Based on these results, we develop practical algorithms, Residual Actor-Critic and Stackelberg Actor-Critic, for estimating the correction between AC and PG and use these to modify the standard AC algorithm. Experiments on popular tabular and continuous environments show the proposed corrections can improve both the sample efficiency and final performance of existing AC methods.
****
Toward Understanding the Feature Learning Process of Self-supervised Contrastive Learning
Zixin Wen, Yuanzhi Li
We formally study how contrastive learning learns the feature representations for neural networks by investigating its feature learning process. We consider the case where our data are comprised of two types of features: the sparse features which we want to learn from, and the dense features we want to get rid of. Theoretically, we prove that contrastive learning using ReLU networks provably learns the desired features if proper augmentations are adopted. We present an underlying principle called feature decoupling to explain the effects of augmentations, where we theoretically characterize how augmentations can reduce the correlations of dense features between positive samples while keeping the correlations of sparse features intact, thereby forcing the neural networks to learn from the self-supervision of sparse features. Empirically, we verified that the feature decoupling principle matches the underlying mechanism of contrastive learning in practice.
****
Keyframe-Focused Visual Imitation Learning
Chuan Wen, Jierui Lin, Jianing Qian, Yang Gao, Dinesh Jayaraman
Imitation learning trains control policies by mimicking pre-recorded expert demonstrations. In partially observable settings, imitation policies must rely on observation histories, but many seemingly paradoxical results show better performance for policies that only access the most recent observation. Recent solutions ranging from causal graph learning to deep information bottlenecks have shown promising results, but failed to scale to realistic settings such as visual imitation. We propose a solution that outperforms these prior approaches by upweighting demonstration keyframes corresponding to expert action changepoints. This simple approach easily scales to complex visual imitation settings. Our experimental results demonstrate consistent performance improvements over all baselines on image-based Gym MuJoCo continuous control tasks. Finally, on the CARLA photorealistic vision-based urban driving simulator, we resolve a long-standing issue in behavioral cloning for driving by demonstrating effective imitation from observation histories. Supplementary materials and code at: \url{https://tinyurl.com/imitation-keyframes}.
****
Learning de-identified representations of prosody from raw audio
Jack Weston, Raphael Lenain, Udeepa Meepegama, Emil Fristed
We propose a method for learning de-identified prosody representations from raw audio using a contrastive self-supervised signal. Whereas prior work has relied on conditioning models with bottlenecks, we introduce a set of inductive biases that exploit the natural structure of prosody to minimize timbral information and decouple prosody from speaker representations. Despite aggressive downsampling of the input and having no access to linguistic information, our model performs comparably to state-of-the-art speech representations on DAMMP, a new benchmark we introduce for spoken language understanding. We use minimum description length probing to show that our representations have selectively learned the subcomponents of non-timbral prosody, and that the product quantizer naturally disentangles them without using bottlenecks. We derive an information-theoretic definition of speech de-identifiability and use it to demonstrate that our prosody representations are less identifiable than the other speech representations.
****
Solving Inverse Problems with a Flow-based Noise Model

Jay Whang, Qi Lei, Alex Dimakis
We study image inverse problems with a normalizing flow prior. Our formulation views the solution as the maximum a posteriori estimate of the image conditioned on the measurements. This formulation allows us to use noise models with arbitrary dependencies as well as non-linear forward operators. We empirically validate the efficacy of our method on various inverse problems, including compressed sensing with quantized measurements and denoising with highly structured noise patterns. We also present initial theoretical recovery guarantees for solving inverse problems with a flow prior.
****

Composing Normalizing Flows for Inverse Problems
Jay Whang, Erik Lindgren, Alex Dimakis
Given an inverse problem with a normalizing flow prior, we wish to estimate the distribution of the underlying signal conditioned on the observations. We approach this problem as a task of conditional inference on the pre-trained unconditional flow model. We first establish that this is computationally hard for a large class of flow models. Motivated by this, we propose a framework for approximate inference that estimates the target conditional as a composition of two flow models. This formulation leads to a stable variational inference training procedure that avoids adversarial training. Our method is evaluated on a variety of inverse problems and is shown to produce high-quality samples with uncertainty quantification. We further demonstrate that our approach can be amortized for zero-shot inference.
****

Which transformer architecture fits my data? A vocabulary bottleneck in self-attention
Noam Wies, Yoav Levine, Daniel Jannai, Amnon Shashua
After their successful debut in natural language processing, Transformer architectures are now becoming the de-facto standard in many domains. An obstacle for their deployment over new modalities is the architectural configuration: the optimal depth-to-width ratio has been shown to dramatically vary across data types (i.e., 10x larger over images than over language). We theoretically predict the existence of an embedding rank bottleneck that limits the contribution of self-attention width to the Transformer expressivity. We thus directly tie the input vocabulary size and rank to the optimal depth-to-width ratio, since a small vocabulary size or rank dictates an added advantage of depth over width. We empirically demonstrate the existence of this bottleneck and its implications on the depth-to-width interplay of Transformer architectures, linking the architecture variability across domains to the often glossed-over usage of different vocabulary sizes or embedding ranks in different domains. As an additional benefit, our rank bottlenecking framework allows us to identify size redundancies of 25%-50% in leading NLP models such as ALBERT and T5.
****

Prediction-Centric Learning of Independent Cascade Dynamics from Partial Observations
Mateusz Wilinski, Andrey Lokhov
Spreading processes play an increasingly important role in modeling for diffusion networks, information propagation, marketing and opinion setting. We address the problem of learning of a spreading model such that the predictions generated from this model are accurate and could be subsequently used for the optimization, and control of diffusion dynamics. We focus on a challenging setting where full observations of the dynamics are not available, and standard approaches such as maximum likelihood quickly become intractable for large network instances. We introduce a computationally efficient algorithm, based on a scalable dynamic message-passing approach, which is able to learn parameters of the effective spreading model given only limited information on the activation times of nodes in the network. The popular Independent Cascade model is used to illustrate our approach. We show that tractable inference from the learned model generates a better prediction of marginal probabilities compared to the original model. We develop a systematic procedure for learning a mixture of models which further improves th

e prediction quality.
****
Leveraging Language to Learn Program Abstractions and Search Heuristics
Catherine Wong, Kevin M Ellis, Joshua Tenenbaum, Jacob Andreas
Inductive program synthesis, or inferring programs from examples of desired behavior, offers a general paradigm for building interpretable, robust, andgeneralizable machine learning systems. Effective program synthesis depends on two key ingredients: a strong library of functions from which to build programs, and an efficient search strategy for finding programs that solve a given task. We introduce LAPS (Language for Abstraction and Program Search), a technique for using natural language annotations to guide joint learning of libraries and neurally-guided search models for synthesis. When integrated into a state-of-the-art library learning system (DreamCoder), LAPS produces higher-quality libraries and improves search efficiency and generalization on three domains {–} string editing, image composition, and abstract reasoning about scenes {–} even when no natural language hints are available at test time.
****
Leveraging Sparse Linear Layers for Debuggable Deep Networks
Eric Wong, Shibani Santurkar, Aleksander Madry
We show how fitting sparse linear models over learned deep feature representations can lead to more debuggable neural networks. These networks remain highly accurate while also being more amenable to human interpretation, as we demonstrate quantitatively and via human experiments. We further illustrate how the resulting sparse explanations can help to identify spurious correlations, explain misclassifications, and diagnose model biases in vision and language tasks.
****
Learning Neural Network Subspaces
Mitchell Wortsman, Maxwell C Horton, Carlos Guestrin, Ali Farhadi, Mohammad Rastegari
Recent observations have advanced our understanding of the neural network optimization landscape, revealing the existence of (1) paths of high accuracy containing diverse solutions and (2) wider minima offering improved performance. Previous methods observing diverse paths require multiple training runs. In contrast we aim to leverage both property (1) and (2) with a single method and in a single training run. With a similar computational cost as training one model, we learn lines, curves, and simplexes of high-accuracy neural networks. These neural network subspaces contain diverse solutions that can be ensembled, approaching the ensemble performance of independently trained networks without the training cost. Moreover, using the subspace midpoint boosts accuracy, calibration, and robustness to label noise, outperforming Stochastic Weight Averaging.
****
Conjugate Energy-Based Models
Hao Wu, Babak Esmaeili, Michael Wick, Jean-Baptiste Tristan, Jan-Willem Van De Meent
In this paper, we propose conjugate energy-based models (CEBMs), a new class of energy-based models that define a joint density over data and latent variables. The joint density of a CEBM decomposes into an intractable distribution over data and a tractable posterior over latent variables. CEBMs have similar use cases as variational autoencoders, in the sense that they learn an unsupervised mapping from data to latent variables. However, these models omit a generator network, which allows them to learn more flexible notions of similarity between data points. Our experiments demonstrate that conjugate EBMs achieve competitive results in terms of image modelling, predictive power of latent space, and out-of-domain detection on a variety of datasets.
****
Making Paper Reviewing Robust to Bid Manipulation Attacks
Ruihan Wu, Chuan Guo, Felix Wu, Rahul Kidambi, Laurens Van Der Maaten, Kilian Weinberger
Most computer science conferences rely on paper bidding to assign reviewers to papers. Although paper bidding enables high-quality assignments in days of unprec

edented submission numbers, it also opens the door for dishonest reviewers to ad versarially influence paper reviewing assignments. Anecdotal evidence suggests t hat some reviewers bid on papers by "friends" or colluding authors, even though these papers are outside their area of expertise, and recommend them for accepta nce without considering the merit of the work. In this paper, we study the effic acy of such bid manipulation attacks and find that, indeed, they can jeopardize the integrity of the review process. We develop a novel approach for paper biddi ng and assignment that is much more robust against such attacks. We show empiric ally that our approach provides robustness even when dishonest reviewers collude , have full knowledge of the assignment system’s internal workings, and have acc ess to the system’s inputs. In addition to being more robust, the quality of our paper review assignments is comparable to that of current, non-robust assignmen t approaches.
****
LIME: Learning Inductive Bias for Primitives of Mathematical Reasoning
Yuhuai Wu, Markus N Rabe, Wenda Li, Jimmy Ba, Roger B Grosse, Christian Szegedy
While designing inductive bias in neural architectures has been widely studied, we hypothesize that transformer networks are flexible enough to learn inductive bias from suitable generic tasks. Here, we replace architecture engineering by e ncoding inductive bias in the form of datasets. Inspired by Peirce’s view that d eduction, induction, and abduction are the primitives of reasoning, we design th ree synthetic tasks that are intended to require the model to have these three a bilities. We specifically design these tasks to be synthetic and devoid of mathe matical knowledge to ensure that only the fundamental reasoning biases can be le arned from these tasks. This defines a new pre-training methodology called "LIME " (Learning Inductive bias for Mathematical rEasoning). Models trained with LIME significantly outperform vanilla transformers on four very different large math ematical reasoning benchmarks. Unlike dominating the computation cost as traditi onal pre-training approaches, LIME requires only a small fraction of the computa tion cost of the typical downstream task. The code for generating LIME tasks is available at https://github.com/tonywu95/LIME.
****
ChaCha for Online AutoML
Qingyun Wu, Chi Wang, John Langford, Paul Mineiro, Marco Rossi
We propose the ChaCha (Champion-Challengers) algorithm for making an online choi ce of hyperparameters in online learning settings. ChaCha handles the process of determining a champion and scheduling a set of ‘live’ challengers over time bas ed on sample complexity bounds. It is guaranteed to have sublinear regret after the optimal configuration is added into consideration by an application-dependen t oracle based on the champions. Empirically, we show that ChaCha provides good performance across a wide array of datasets when optimizing over featurization a nd hyperparameter decisions.
****
Temporally Correlated Task Scheduling for Sequence Learning
Xueqing Wu, Lewen Wang, Yingce Xia, Weiqing Liu, Lijun Wu, Shufang Xie, Tao Qin, Tie-Yan Liu
Sequence learning has attracted much research attention from the machine learnin g community in recent years. In many applications, a sequence learning task is u sually associated with multiple temporally correlated auxiliary tasks, which are different in terms of how much input information to use or which future step to predict. For example, (i) in simultaneous machine translation, one can conduct translation under different latency (i.e., how many input words to read/wait bef ore translation); (ii) in stock trend forecasting, one can predict the price of a stock in different future days (e.g., tomorrow, the day after tomorrow). While it is clear that those temporally correlated tasks can help each other, there i s a very limited exploration on how to better leverage multiple auxiliary tasks to boost the performance of the main task. In this work, we introduce a learnabl e scheduler to sequence learning, which can adaptively select auxiliary tasks fo r training depending on the model status and the current training data. The sche duler and the model for the main task are jointly trained through bi-level optim

ization. Experiments show that our method significantly improves the performance of simultaneous machine translation and stock trend forecasting.
****
Class2Simi: A Noise Reduction Perspective on Learning with Noisy Labels
Songhua Wu, Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Nannan Wang, Haifeng Liu, Gang Niu
Learning with noisy labels has attracted a lot of attention in recent years, where the mainstream approaches are in \emph{pointwise} manners. Meanwhile, \emph{pairwise} manners have shown great potential in supervised metric learning and unsupervised contrastive learning. Thus, a natural question is raised: does learning in a pairwise manner \emph{mitigate} label noise? To give an affirmative answer, in this paper, we propose a framework called \emph{Class2Simi}: it transforms data points with noisy \emph{class labels} to data pairs with noisy \emph{similarity labels}, where a similarity label denotes whether a pair shares the class label or not. Through this transformation, the \emph{reduction of the noise rate} is theoretically guaranteed, and hence it is in principle easier to handle noisy similarity labels. Amazingly, DNNs that predict the \emph{clean} class labels can be trained from noisy data pairs if they are first pretrained from noisy data points. Class2Simi is \emph{computationally efficient} because not only this transformation is on-the-fly in mini-batches, but also it just changes loss computation on top of model prediction into a pairwise manner. Its effectiveness is verified by extensive experiments.
****
On Reinforcement Learning with Adversarial Corruption and Its Application to Block MDP
Tianhao Wu, Yunchang Yang, Simon Du, Liwei Wang
We study reinforcement learning (RL) in episodic tabular MDPs with adversarial corruptions, where some episodes can be adversarially corrupted. When the total number of corrupted episodes is known, we propose an algorithm, Corruption Robust Monotonic Value Propagation (\textsf{CR-MVP}), which achieves a regret bound of $\tilde{O}\left(\left(\sqrt{SAK}+S^2A+CSA)\right)\polylog(H)\right)$, where S is the number of states, A is the number of actions, H is the planning horizon, K is the number of episodes, and C is the corruption level. We also provide a corresponding lower bound, which indicates that our upper bound is tight. Finally, as an application, we study RL with rich observations in the block MDP model. We provide the first algorithm that achieves a $\sqrt{K}$-type regret in this setting and is computationally efficient.
****
Generative Video Transformer: Can Objects be the Words?
Yi-Fu Wu, Jaesik Yoon, Sungjin Ahn
Transformers have been successful for many natural language processing tasks. However, applying transformers to the video domain for tasks such as long-term video generation and scene understanding has remained elusive due to the high computational complexity and the lack of natural tokenization. In this paper, we propose the ObjectCentric Video Transformer (OCVT) which utilizes an object-centric approach for decomposing scenes into tokens suitable for use in a generative video transformer. By factoring the video into objects, our fully unsupervised model is able to learn complex spatio-temporal dynamics of multiple interacting objects in a scene and generate future frames of the video. Our model is also significantly more memory-efficient than pixel-based models and thus able to train on videos of length up to 70 frames with a single 48GB GPU. We compare our model with previous RNN-based approaches as well as other possible video transformer baselines. We demonstrate OCVT performs well when compared to baselines in generating future frames. OCVT also develops useful representations for video reasoning, achieving start-of-the-art performance on the CATER task.
****
Uncertainty Weighted Actor-Critic for Offline Reinforcement Learning
Yue Wu, Shuangfei Zhai, Nitish Srivastava, Joshua M Susskind, Jian Zhang, Ruslan Salakhutdinov, Hanlin Goh
Offline Reinforcement Learning promises to learn effective policies from previou

sly-collected, static datasets without the need for exploration. However, existing Q-learning and actor-critic based off-policy RL algorithms fail when bootstrapping from out-of-distribution (OOD) actions or states. We hypothesize that a key missing ingredient from the existing methods is a proper treatment of uncertainty in the offline setting. We propose Uncertainty Weighted Actor-Critic (UWAC), an algorithm that detects OOD state-action pairs and down-weights their contribution in the training objectives accordingly. Implementation-wise, we adopt a practical and effective dropout-based uncertainty estimation method that introduces very little overhead over existing RL algorithms. Empirically, we observe that UWAC substantially improves model stability during training. In addition, UWAC out-performs existing offline RL methods on a variety of competitive tasks, and achieves significant performance gains over the state-of-the-art baseline on datasets with sparse demonstrations collected from human experts.
****
Towards Open-World Recommendation: An Inductive Model-based Collaborative Filtering Approach
Qitian Wu, Hengrui Zhang, Xiaofeng Gao, Junchi Yan, Hongyuan Zha
Recommendation models can effectively estimate underlying user interests and predict one's future behaviors by factorizing an observed user-item rating matrix into products of two sets of latent factors. However, the user-specific embedding factors can only be learned in a transductive way, making it difficult to handle new users on-the-fly. In this paper, we propose an inductive collaborative filtering framework that contains two representation models. The first model follows conventional matrix factorization which factorizes a group of key users' rating matrix to obtain meta latents. The second model resorts to attention-based structure learning that estimates hidden relations from query to key users and learns to leverage meta latents to inductively compute embeddings for query users via a neural message passing. Our model enables inductive representation learning for users and meanwhile guarantees equivalent representation capacity as matrix factorization. Experiments demonstrate that our model achieves promising results for recommendation on few-shot users with limited training ratings and new unseen users which are commonly encountered in open-world recommender systems.
****
Data-efficient Hindsight Off-policy Option Learning
Markus Wulfmeier, Dushyant Rao, Roland Hafner, Thomas Lampe, Abbas Abdolmaleki, Tim Hertweck, Michael Neunert, Dhruva Tirumala, Noah Siegel, Nicolas Heess, Martin Riedmiller
We introduce Hindsight Off-policy Options (HO2), a data-efficient option learning algorithm. Given any trajectory, HO2 infers likely option choices and backpropagates through the dynamic programming inference procedure to robustly train all policy components off-policy and end-to-end. The approach outperforms existing option learning methods on common benchmarks. To better understand the option framework and disentangle benefits from both temporal and action abstraction, we evaluate ablations with flat policies and mixture policies with comparable optimization. The results highlight the importance of both types of abstraction as well as off-policy training and trust-region constraints, particularly in challenging, simulated 3D robot manipulation tasks from raw pixel inputs. Finally, we intuitively adapt the inference step to investigate the effect of increased temporal abstraction on training with pre-trained options and from scratch.
****
A Bit More Bayesian: Domain-Invariant Learning with Uncertainty
Zehao Xiao, Jiayi Shen, Xiantong Zhen, Ling Shao, Cees Snoek
Domain generalization is challenging due to the domain shift and the uncertainty caused by the inaccessibility of target domain data. In this paper, we address both challenges with a probabilistic framework based on variational Bayesian inference, by incorporating uncertainty into neural network weights. We couple domain invariance in a probabilistic formula with the variational Bayesian inference. This enables us to explore domain-invariant learning in a principled way. Specifically, we derive domain-invariant representations and classifiers, which are jointly established in a two-layer Bayesian neural network. We empirically demon

strate the effectiveness of our proposal on four widely used cross-domain visual recognition benchmarks. Ablation studies validate the synergistic benefits of our Bayesian treatment when jointly learning domain-invariant representations and classifiers for domain generalization. Further, our method consistently delivers state-of-the-art mean accuracy on all benchmarks.
****
On the Optimality of Batch Policy Optimization Algorithms
Chenjun Xiao, Yifan Wu, Jincheng Mei, Bo Dai, Tor Lattimore, Lihong Li, Csaba Szepesvari, Dale Schuurmans
Batch policy optimization considers leveraging existing data for policy construction before interacting with an environment. Although interest in this problem has grown significantly in recent years, its theoretical foundations remain under-developed. To advance the understanding of this problem, we provide three results that characterize the limits and possibilities of batch policy optimization in the finite-armed stochastic bandit setting. First, we introduce a class of confidence-adjusted index algorithms that unifies optimistic and pessimistic principles in a common framework, which enables a general analysis. For this family, we show that any confidence-adjusted index algorithm is minimax optimal, whether it be optimistic, pessimistic or neutral. Our analysis reveals that instance-dependent optimality, commonly used to establish optimality of on-line stochastic bandit algorithms, cannot be achieved by any algorithm in the batch setting. In particular, for any algorithm that performs optimally in some environment, there exists another environment where the same algorithm suffers arbitrarily larger regret. Therefore, to establish a framework for distinguishing algorithms, we introduce a new weighted-minimax criterion that considers the inherent difficulty of optimal value prediction. We demonstrate how this criterion can be used to justify commonly used pessimistic principles for batch policy optimization.
****
CRFL: Certifiably Robust Federated Learning against Backdoor Attacks
Chulin Xie, Minghao Chen, Pin-Yu Chen, Bo Li
Federated Learning (FL) as a distributed learning paradigm that aggregates information from diverse clients to train a shared global model, has demonstrated great success. However, malicious clients can perform poisoning attacks and model replacement to introduce backdoors into the trained global model. Although there have been intensive studies designing robust aggregation methods and empirical robust federated training protocols against backdoors, existing approaches lack robustness certification. This paper provides the first general framework, Certifiably Robust Federated Learning (CRFL), to train certifiably robust FL models against backdoors. Our method exploits clipping and smoothing on model parameters to control the global model smoothness, which yields a sample-wise robustness certification on backdoors with limited magnitude. Our certification also specifies the relation to federated learning parameters, such as poisoning ratio on instance level, number of attackers, and training iterations. Practically, we conduct comprehensive experiments across a range of federated datasets, and provide the first benchmark for certified robustness against backdoor attacks in federated learning. Our code is publicaly available at https://github.com/AI-secure/CRFL.
****
RNNRepair: Automatic RNN Repair via Model-based Analysis
Xiaofei Xie, Wenbo Guo, Lei Ma, Wei Le, Jian Wang, Lingjun Zhou, Yang Liu, Xinyu Xing
Deep neural networks are vulnerable to adversarial attacks. Due to their black-box nature, it is rather challenging to interpret and properly repair these incorrect behaviors. This paper focuses on interpreting and repairing the incorrect behaviors of Recurrent Neural Networks (RNNs). We propose a lightweight model-based approach (RNNRepair) to help understand and repair incorrect behaviors of an RNN. Specifically, we build an influence model to characterize the stateful and statistical behaviors of an RNN over all the training data and to perform the influence analysis for the errors. Compared with the existing techniques on influence function, our method can efficiently estimate the influence of existing or newly added training samples for a given prediction at both sample level and segm

entation level. Our empirical evaluation shows that the proposed influence model is able to extract accurate and understandable features. Based on the influence model, our proposed technique could effectively infer the influential instances from not only an entire testing sequence but also a segment within that sequence. Moreover, with the sample-level and segment-level influence relations, RNNRepair could further remediate two types of incorrect predictions at the sample level and segment level.
****
Deep Reinforcement Learning amidst Continual Structured Non-Stationarity
Annie Xie, James Harrison, Chelsea Finn
As humans, our goals and our environment are persistently changing throughout our lifetime based on our experiences, actions, and internal and external drives. In contrast, typical reinforcement learning problem set-ups consider decision processes that are stationary across episodes. Can we develop reinforcement learning algorithms that can cope with the persistent change in the former, more realistic problem settings? While on-policy algorithms such as policy gradients in principle can be extended to non-stationary settings, the same cannot be said for more efficient off-policy algorithms that replay past experiences when learning. In this work, we formalize this problem setting, and draw upon ideas from the online learning and probabilistic inference literature to derive an off-policy RL algorithm that can reason about and tackle such lifelong non-stationarity. Our method leverages latent variable models to learn a representation of the environment from current and past experiences, and performs off-policy RL with this representation. We further introduce several simulation environments that exhibit lifelong non-stationarity, and empirically find that our approach substantially outperforms approaches that do not reason about environment shift.
****
Batch Value-function Approximation with Only Realizability
Tengyang Xie, Nan Jiang
We make progress in a long-standing problem of batch reinforcement learning (RL): learning Q* from an exploratory and polynomial-sized dataset, using a realizable and otherwise arbitrary function class. In fact, all existing algorithms demand function-approximation assumptions stronger than realizability, and the mounting negative evidence has led to a conjecture that sample-efficient learning is impossible in this setting (Chen & Jiang, 2019). Our algorithm, BVFT, breaks the hardness conjecture (albeit under a stronger notion of exploratory data) via a tournament procedure that reduces the learning problem to pairwise comparison, and solves the latter with the help of a state-action-space partition constructed from the compared functions. We also discuss how BVFT can be applied to model selection among other extensions and open problems.
****
Interaction-Grounded Learning
Tengyang Xie, John Langford, Paul Mineiro, Ida Momennejad
Consider a prosthetic arm, learning to adapt to its user’s control signals. We propose \emph{Interaction-Grounded Learning} for this novel setting, in which a learner’s goal is to interact with the environment with no grounding or explicit reward to optimize its policies. Such a problem evades common RL solutions which require an explicit reward. The learning agent observes a multidimensional \emph{context vector}, takes an \emph{action}, and then observes a multidimensional \emph{feedback vector}. This multidimensional feedback vector has \emph{no} explicit reward information. In order to succeed, the algorithm must learn how to evaluate the feedback vector to discover a latent reward signal, with which it can ground its policies without supervision. We show that in an Interaction-Grounded Learning setting, with certain natural assumptions, a learner can discover the latent reward and ground its policy for successful interaction. We provide theoretical guarantees and a proof-of-concept empirical evaluation to demonstrate the effectiveness of our proposed approach.
****
Composed Fine-Tuning: Freezing Pre-Trained Denoising Autoencoders for Improved Generalization

Sang Michael Xie, Tengyu Ma, Percy Liang
We focus on prediction problems with structured outputs that are subject to output validity constraints, e.g. pseudocode-to-code translation where the code must compile. While labeled input-output pairs are expensive to obtain, "unlabeled" outputs, i.e. outputs without corresponding inputs, are freely available (e.g. code on GitHub) and provide information about output validity. Pre-training captures this structure by training a denoiser to denoise corrupted versions of unlabeled outputs. We first show that standard fine-tuning after pre-training destroys some of this structure. We then propose composed fine-tuning, which trains a predictor composed with the pre-trained denoiser. Importantly, the denoiser is fixed to preserve output structure. Like standard fine-tuning, the predictor is also initialized with the pre-trained denoiser. We prove for two-layer ReLU networks that composed fine-tuning significantly reduces the complexity of the predictor, thus improving generalization. Empirically, we show that composed fine-tuning improves over standard fine-tuning on two pseudocode-to-code translation datasets (3% and 6% relative). The improvement is magnified on out-of-distribution (OOD) examples (4% and 25% relative), suggesting that reducing predictor complexity improves OOD extrapolation.
****
Learning While Playing in Mean-Field Games: Convergence and Optimality
Qiaomin Xie, Zhuoran Yang, Zhaoran Wang, Andreea Minca
We study reinforcement learning in mean-field games. To achieve the Nash equilibrium, which consists of a policy and a mean-field state, existing algorithms require obtaining the optimal policy while fixing any mean-field state. In practice, however, the policy and the mean-field state evolve simultaneously, as each agent is learning while playing. To bridge such a gap, we propose a fictitious play algorithm, which alternatively updates the policy (learning) and the mean-field state (playing) by one step of policy optimization and gradient descent, respectively. Despite the nonstationarity induced by such an alternating scheme, we prove that the proposed algorithm converges to the Nash equilibrium with an explicit convergence rate. To the best of our knowledge, it is the first provably efficient algorithm that achieves learning while playing via alternating updates.
****
Positive-Negative Momentum: Manipulating Stochastic Gradient Noise to Improve Generalization
Zeke Xie, Li Yuan, Zhanxing Zhu, Masashi Sugiyama
It is well-known that stochastic gradient noise (SGN) acts as implicit regularization for deep learning and is essentially important for both optimization and generalization of deep networks. Some works attempted to artificially simulate SGN by injecting random noise to improve deep learning. However, it turned out that the injected simple random noise cannot work as well as SGN, which is anisotropic and parameter-dependent. For simulating SGN at low computational costs and without changing the learning rate or batch size, we propose the Positive-Negative Momentum (PNM) approach that is a powerful alternative to conventional Momentum in classic optimizers. The introduced PNM method maintains two approximate independent momentum terms. Then, we can control the magnitude of SGN explicitly by adjusting the momentum difference. We theoretically prove the convergence guarantee and the generalization advantage of PNM over Stochastic Gradient Descent (SGD). By incorporating PNM into the two conventional optimizers, SGD with Momentum and Adam, our extensive experiments empirically verified the significant advantage of the PNM-based variants over the corresponding conventional Momentum-based optimizers. Code: \url{https://github.com/zeke-xie/Positive-Negative-Momentum}.
****
A Hybrid Variance-Reduced Method for Decentralized Stochastic Non-Convex Optimization
Ran Xin, Usman Khan, Soummya Kar
This paper considers decentralized stochastic optimization over a network of n nodes, where each node possesses a smooth non-convex local cost function and the goal of the networked nodes is to find an ϵ-accurate first-order stat

ionary point of the sum of the local costs. We focus on an online setting, where each node accesses its local cost only by means of a stochastic first-order oracle that returns a noisy version of the exact gradient. In this context, we propose a novel single-loop decentralized hybrid variance-reduced stochastic gradient method, called GT-HSGD, that outperforms the existing approaches in terms of both the oracle complexity and practical implementation. The GT-HSGD algorithm implements specialized local hybrid stochastic gradient estimators that are fused over the network to track the global gradient. Remarkably, GT-HSGD achieves a network topology-independent oracle complexity of $O(n^{-1}\epsilon^{-3})$ when the required error tolerance ϵ is small enough, leading to a linear speedup with respect to the centralized optimal online variance-reduced approaches that operate on a single node. Numerical experiments are provided to illustrate our main technical results.
****
Explore Visual Concept Formation for Image Classification
Shengzhou Xiong, Yihua Tan, Guoyou Wang
Human beings acquire the ability of image classification through visual concept learning, in which the process of concept formation involves intertwined searches of common properties and concept descriptions. However, in most image classification algorithms using deep convolutional neural network (ConvNet), the representation space is constructed under the premise that concept descriptions are fixed as one-hot codes, which limits the mining of properties and the ability of identifying unseen samples. Inspired by this, we propose a learning strategy of visual concept formation (LSOVCF) based on the ConvNet, in which the two intertwined parts of concept formation, i.e. feature extraction and concept description, are learned together. First, LSOVCF takes sample response in the last layer of ConvNet to induct concept description being assumed as Gaussian distribution, which is part of the training process. Second, the exploration and experience loss is designed for optimization, which adopts experience cache pool to speed up convergence. Experiments show that LSOVCF improves the ability of identifying unseen samples on cifar10, STL10, flower17 and ImageNet based on several backbones, from the classic VGG to the SOTA Ghostnet. The code is available at \url{https://github.com/elvintanhust/LSOVCF}.
****
CRPO: A New Approach for Safe Reinforcement Learning with Convergence Guarantee
Tengyu Xu, Yingbin Liang, Guanghui Lan
In safe reinforcement learning (SRL) problems, an agent explores the environment to maximize an expected total reward and meanwhile avoids violation of certain constraints on a number of expected total costs. In general, such SRL problems have nonconvex objective functions subject to multiple nonconvex constraints, and hence are very challenging to solve, particularly to provide a globally optimal policy. Many popular SRL algorithms adopt a primal-dual structure which utilizes the updating of dual variables for satisfying the constraints. In contrast, we propose a primal approach, called constraint-rectified policy optimization (CRPO), which updates the policy alternatingly between objective improvement and constraint satisfaction. CRPO provides a primal-type algorithmic framework to solve SRL problems, where each policy update can take any variant of policy optimization step. To demonstrate the theoretical performance of CRPO, we adopt natural policy gradient (NPG) for each policy update step and show that CRPO achieves an $\mathcal{O}(1/\sqrt{T})$ convergence rate to the global optimal policy in the constrained policy set and an $\mathcal{O}(1/\sqrt{T})$ error bound on constraint satisfaction. This is the first finite-time analysis of primal SRL algorithms with global optimality guarantee. Our empirical results demonstrate that CRPO can outperform the existing primal-dual baseline algorithms significantly.
****
To be Robust or to be Fair: Towards Fairness in Adversarial Training
Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, Jiliang Tang
Adversarial training algorithms have been proved to be reliable to improve machine learning models’ robustness against adversarial examples. However, we find that adversarial training algorithms tend to introduce severe disparity of accurac

y and robustness between different groups of data. For instance, PGD adversarial ly trained ResNet18 model on CIFAR-10 has 93% clean accuracy and 67% PGD l_infty -8 adversarial accuracy on the class ”automobile” but only 65% and 17% on class ”cat”. This phenomenon happens in balanced datasets and does not exist in natura lly trained models when only using clean samples. In this work, we empirically a nd theoretically show that this phenomenon can generally happen under adversaria l training algorithms which minimize DNN models’ robust errors. Motivated by the se findings, we propose a Fair-Robust-Learning (FRL) framework to mitigate this unfairness problem when doing adversarial defenses and experimental results vali date the effectiveness of FRL.
****
Interpretable Stein Goodness-of-fit Tests on Riemannian Manifold
Wenkai Xu, Takeru Matsuda
In many applications, we encounter data on Riemannian manifolds such as torus an d rotation groups. Standard statistical procedures for multivariate data are not applicable to such data. In this study, we develop goodness-of-fit testing and interpretable model criticism methods for general distributions on Riemannian ma nifolds, including those with an intractable normalization constant. The propose d methods are based on extensions of kernel Stein discrepancy, which are derived from Stein operators on Riemannian manifolds. We discuss the connections betwee n the proposed tests with existing ones and provide a theoretical analysis of th eir asymptotic Bahadur efficiency. Simulation results and real data applications show the validity and usefulness of the proposed methods.
****
Rethinking Neural vs. Matrix-Factorization Collaborative Filtering: the Theoreti cal Perspectives
Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, Kannan Achan
The recent work by Rendle et al. (2020), based on empirical observations, argues that matrix-factorization collaborative filtering (MCF) compares favorably to n eural collaborative filtering (NCF), and conjectures the dot product’s superiori ty over the feed-forward neural network as similarity function. In this paper, w e address the comparison rigorously by answering the following questions: 1. wha t is the limiting expressivity of each model; 2. under the practical gradient de scent, to which solution does each optimization path converge; 3. how would the models generalize under the inductive and transductive learning setting. Our res ults highlight the similar expressivity for the overparameterized NCF and MCF as kernelized predictors, and reveal the relation between their optimization paths . We further show their different generalization behaviors, where MCF and NCF ex perience specific tradeoff and comparison in the transductive and inductive coll aborative filtering setting. Lastly, by showing a novel generalization result, w e reveal the critical role of correcting exposure bias for model evaluation in t he inductive setting. Our results explain some of the previously observed confli cts, and we provide synthetic and real-data experiments to shed further insights to this topic.
****
Dash: Semi-Supervised Learning with Dynamic Thresholding
Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, Rong Jin
While semi-supervised learning (SSL) has received tremendous attentions in many machine learning tasks due to its successful use of unlabeled data, existing SSL algorithms use either all unlabeled examples or the unlabeled examples with a f ixed high-confidence prediction during the training progress. However, it is pos sible that too many correct/wrong pseudo labeled examples are eliminated/selecte d. In this work we develop a simple yet powerful framework, whose key idea is to select a subset of training examples from the unlabeled data when performing ex isting SSL methods so that only the unlabeled examples with pseudo labels relate d to the labeled data will be used to train models. The selection is performed a t each updating iteration by only keeping the examples whose losses are smaller than a given threshold that is dynamically adjusted through the iteration. Our p roposed approach, Dash, enjoys its adaptivity in terms of unlabeled data selecti on and its theoretical guarantee. Specifically, we theoretically establish the c

onvergence rate of Dash from the view of non-convex optimization. Finally, we em pirically demonstrate the effectiveness of the proposed method in comparison wit h state-of-the-art over benchmarks.
****
An End-to-End Framework for Molecular Conformation Generation via Bilevel Progra mming
Minkai Xu, Wujie Wang, Shitong Luo, Chence Shi, Yoshua Bengio, Rafael Gomez-Bomb arelli, Jian Tang
Predicting molecular conformations (or 3D structures) from molecular graphs is a fundamental problem in many applications. Most existing approaches are usually divided into two steps by first predicting the distances between atoms and then generating a 3D structure through optimizing a distance geometry problem. Howeve r, the distances predicted with such two-stage approaches may not be able to con sistently preserve the geometry of local atomic neighborhoods, making the genera ted structures unsatisfying. In this paper, we propose an end-to-end solution fo r molecular conformation prediction called ConfVAE based on the conditional vari ational autoencoder framework. Specifically, the molecular graph is first encode d in a latent space, and then the 3D structures are generated by solving a princ ipled bilevel optimization program. Extensive experiments on several benchmark d ata sets prove the effectiveness of our proposed approach over existing state-of -the-art approaches. Code is available at \url{https://github.com/MinkaiXu/ConfV AE-ICML21}.
****
Self-supervised Graph-level Representation Learning with Local and Global Struct ure
Minghao Xu, Hang Wang, Bingbing Ni, Hongyu Guo, Jian Tang
This paper studies unsupervised/self-supervised whole-graph representation learn ing, which is critical in many tasks such as molecule properties prediction in d rug and material discovery. Existing methods mainly focus on preserving the loca l similarity structure between different graph instances but fail to discover th e global semantic structure of the entire data set. In this paper, we propose a unified framework called Local-instance and Global-semantic Learning (GraphLoG) for self-supervised whole-graph representation learning. Specifically, besides p reserving the local similarities, GraphLoG introduces the hierarchical prototype s to capture the global semantic clusters. An efficient online expectation-maxim ization (EM) algorithm is further developed for learning the model. We evaluate GraphLoG by pre-training it on massive unlabeled graphs followed by fine-tuning on downstream tasks. Extensive experiments on both chemical and biological bench mark data sets demonstrate the effectiveness of the proposed approach.
****
Conformal prediction interval for dynamic time-series
Chen Xu, Yao Xie
We develop a method to construct distribution-free prediction intervals for dyna mic time-series, called \Verb|EnbPI| that wraps around any bootstrap ensemble es timator to construct sequential prediction intervals. \Verb|EnbPI| is closely re lated to the conformal prediction (CP) framework but does not require data excha ngeability. Theoretically, these intervals attain finite-sample, \textit{approxi mately valid} marginal coverage for broad classes of regression functions and ti me-series with strongly mixing stochastic errors. Computationally, \Verb|EnbPI| avoids overfitting and requires neither data-splitting nor training multiple ens emble estimators; it efficiently aggregates bootstrap estimators that have been trained. In general, \Verb|EnbPI| is easy to implement, scalable to producing ar bitrarily many prediction intervals sequentially, and well-suited to a wide rang e of regression functions. We perform extensive real-data analyses to demonstrat e its effectiveness.
****
Learner-Private Convex Optimization
Jiaming Xu, Kuang Xu, Dana Yang
Convex optimization with feedback is a framework where a learner relies on itera tive queries and feedback to arrive at the minimizer of a convex function. The p

aradigm has gained significant popularity recently thanks to its scalability in large-scale optimization and machine learning. The repeated interactions, however, expose the learner to privacy risks from eavesdropping adversaries that observe the submitted queries. In this paper, we study how to optimally obfuscate the learner's queries in convex optimization with first-order feedback, so that their learned optimal value is provably difficult to estimate for the eavesdropping adversary. We consider two formulations of learner privacy: a Bayesian formulation in which the convex function is drawn randomly, and a minimax formulation in which the function is fixed and the adversary's probability of error is measured with respect to a minimax criterion. We show that, if the learner wants to ensure the probability of the adversary estimating accurately be kept below 1/L, then the overhead in query complexity is additive in L in the minimax formulation, but multiplicative in L in the Bayesian formulation. Compared to existing learner-private sequential learning models with binary feedback, our results apply to the significantly richer family of general convex functions with full-gradient feedback. Our proofs are largely enabled by tools from the theory of Dirichlet processes, as well as more sophisticated lines of analysis aimed at measuring the amount of information leakage under a full-gradient oracle.
****
Doubly Robust Off-Policy Actor-Critic: Convergence and Optimality
Tengyu Xu, Zhuoran Yang, Zhaoran Wang, Yingbin Liang
Designing off-policy reinforcement learning algorithms is typically a very challenging task, because a desirable iteration update often involves an expectation over an on-policy distribution. Prior off-policy actor-critic (AC) algorithms have introduced a new critic that uses the density ratio for adjusting the distribution mismatch in order to stabilize the convergence, but at the cost of potentially introducing high biases due to the estimation errors of both the density ratio and value function. In this paper, we develop a doubly robust off-policy AC (DR-Off-PAC) for discounted MDP, which can take advantage of learned nuisance functions to reduce estimation errors. Moreover, DR-Off-PAC adopts a single timescale structure, in which both actor and critics are updated simultaneously with constant stepsize, and is thus more sample efficient than prior algorithms that adopt either two timescale or nested-loop structure. We study the finite-time convergence rate and characterize the sample complexity for DR-Off-PAC to attain an ϵ-accurate optimal policy. We also show that the overall convergence of DR-Off-PAC is doubly robust to the approximation errors that depend only on the expressive power of approximation functions. To the best of our knowledge, our study establishes the first overall sample complexity analysis for single time-scale off-policy AC algorithm.
****
Optimization of Graph Neural Networks: Implicit Acceleration by Skip Connections and More Depth
Keyulu Xu, Mozhi Zhang, Stefanie Jegelka, Kenji Kawaguchi
Graph Neural Networks (GNNs) have been studied through the lens of expressive power and generalization. However, their optimization properties are less well understood. We take the first step towards analyzing GNN training by studying the gradient dynamics of GNNs. First, we analyze linearized GNNs and prove that despite the non-convexity of training, convergence to a global minimum at a linear rate is guaranteed under mild assumptions that we validate on real-world graphs. Second, we study what may affect the GNNs' training speed. Our results show that the training of GNNs is implicitly accelerated by skip connections, more depth, and/or a good label distribution. Empirical results confirm that our theoretical results for linearized GNNs align with the training behavior of nonlinear GNNs. Our results provide the first theoretical support for the success of GNNs with skip connections in terms of optimization, and suggest that deep GNNs with skip connections would be promising in practice.
****
Group-Sparse Matrix Factorization for Transfer Learning of Word Embeddings
Kan Xu, Xuanyi Zhao, Hamsa Bastani, Osbert Bastani
Sparse regression has recently been applied to enable transfer learning from ver

y limited data. We study an extension of this approach to unsupervised learning—in particular, learning word embeddings from unstructured text corpora using low-rank matrix factorization. Intuitively, when transferring word embeddings to a new domain, we expect that the embeddings change for only a small number of words—e.g., the ones with novel meanings in that domain. We propose a novel group-sparse penalty that exploits this sparsity to perform transfer learning when there is very little text data available in the target domain—e.g., a single article of text. We prove generalization bounds for our algorithm. Furthermore, we empirically evaluate its effectiveness, both in terms of prediction accuracy in downstream tasks as well as in terms of interpretability of the results.
****
KNAS: Green Neural Architecture Search
Jingjing Xu, Liang Zhao, Junyang Lin, Rundong Gao, Xu Sun, Hongxia Yang
Many existing neural architecture search (NAS) solutions rely on downstream training for architecture evaluation, which takes enormous computations. Considering that these computations bring a large carbon footprint, this paper aims to explore a green (namely environmental-friendly) NAS solution that evaluates architectures without training. Intuitively, gradients, induced by the architecture itself, directly decide the convergence and generalization results. It motivates us to propose the gradient kernel hypothesis: Gradients can be used as a coarse-grained proxy of downstream training to evaluate random-initialized networks. To support the hypothesis, we conduct a theoretical analysis and find a practical gradient kernel that has good correlations with training loss and validation performance. According to this hypothesis, we propose a new kernel based architecture search approach KNAS. Experiments show that KNAS achieves competitive results with orders of magnitude faster than “train-then-test” paradigms on image classification tasks. Furthermore, the extremely low search cost enables its wide applications. The searched network also outperforms strong baseline RoBERTA-large on two text classification tasks.
****
Structured Convolutional Kernel Networks for Airline Crew Scheduling
Yassine Yaakoubi, Francois Soumis, Simon Lacoste-Julien
Motivated by the needs from an airline crew scheduling application, we introduce structured convolutional kernel networks (Struct-CKN), which combine CKNs from Mairal et al. (2014) in a structured prediction framework that supports constraints on the outputs. CKNs are a particular kind of convolutional neural networks that approximate a kernel feature map on training data, thus combining properties of deep learning with the non-parametric flexibility of kernel methods. Extending CKNs to structured outputs allows us to obtain useful initial solutions on a flight-connection dataset that can be further refined by an airline crew scheduling solver. More specifically, we use a flight-based network modeled as a general conditional random field capable of incorporating local constraints in the learning process. Our experiments demonstrate that this approach yields significant improvements for the large-scale crew pairing problem (50,000 flights per month) over standard approaches, reducing the solution cost by 17% (a gain of millions of dollars) and the cost of global constraints by 97%.
****
Mediated Uncoupled Learning: Learning Functions without Direct Input-output Correspondences
Ikko Yamane, Junya Honda, Florian Yger, Masashi Sugiyama
Ordinary supervised learning is useful when we have paired training data of input X and output Y. However, such paired data can be difficult to collect in practice. In this paper, we consider the task of predicting Y from X when we have no paired data of them, but we have two separate, independent datasets of X and Y each observed with some mediating variable U, that is, we have two datasets $S_X = \{(X_i, U_i)\}$ and $S_Y = \{(U'_j, Y'_j)\}$. A naive approach is to predict U from X using S_X and then Y from U using S_Y, but we show that this is not statistically consistent. Moreover, predicting U can be more difficult than predicting Y in practice, e.g., when U has higher dimensionality. To circumvent the difficulty, we propose a new method that avoids pred

icting U but directly learns $Y = f(X)$ by training $f(X)$ with S_{X} to pre dict $h(U)$ which is trained with S_{Y} to approximate Y. We prove statistic al consistency and error bounds of our method and experimentally confirm its pra ctical usefulness.
****
EL-Attention: Memory Efficient Lossless Attention for Generation
Yu Yan, Jiusheng Chen, Weizhen Qi, Nikhil Bhendawade, Yeyun Gong, Nan Duan, Ruof ei Zhang
Transformer model with multi-head attention requires caching intermediate result s for efficient inference in generation tasks. However, cache brings new memory-related costs and prevents leveraging larger batch size for faster speed. We pro pose memory-efficient lossless attention (called EL-attention) to address this i ssue. It avoids heavy operations for building multi-head keys and values, cache for them is not needed. EL-attention constructs an ensemble of attention results by expanding query while keeping key and value shared. It produces the same res ult as multi-head attention with less GPU memory and faster inference speed. We conduct extensive experiments on Transformer, BART, and GPT-2 for summarization and question generation tasks. The results show EL-attention speeds up existing models by 1.6x to 5.3x without accuracy loss.
****
Link Prediction with Persistent Homology: An Interactive View
Zuoyu Yan, Tengfei Ma, Liangcai Gao, Zhi Tang, Chao Chen
Link prediction is an important learning task for graph-structured data. In this paper, we propose a novel topological approach to characterize interactions bet ween two nodes. Our topological feature, based on the extended persistent homolo gy, encodes rich structural information regarding the multi-hop paths connecting nodes. Based on this feature, we propose a graph neural network method that out performs state-of-the-arts on different benchmarks. As another contribution, we propose a novel algorithm to more efficiently compute the extended persistence d iagrams for graphs. This algorithm can be generally applied to accelerate many o ther topological methods for graph learning tasks.
****
CATE: Computation-aware Neural Architecture Encoding with Transformers
Shen Yan, Kaiqiang Song, Fei Liu, Mi Zhang
Recent works (White et al., 2020a; Yan et al., 2020) demonstrate the importance of architecture encodings in Neural Architecture Search (NAS). These encodings e ncode either structure or computation information of the neural architectures. C ompared to structure-aware encodings, computation-aware encodings map architectu res with similar accuracies to the same region, which improves the downstream ar chitecture search performance (Zhang et al., 2019; White et al., 2020a). In this work, we introduce a Computation-Aware Transformer-based Encoding method called CATE. Different from existing computation-aware encodings based on fixed transf ormation (e.g. path encoding), CATE employs a pairwise pre-training scheme to le arn computation-aware encodings using Transformers with cross-attention. Such le arned encodings contain dense and contextualized computation information of neur al architectures. We compare CATE with eleven encodings under three major encodi ng-dependent NAS subroutines in both small and large search spaces. Our experime nts show that CATE is beneficial to the downstream search, especially in the lar ge search space. Moreover, the outside search space experiment demonstrates its superior generalization ability beyond the search space on which it was trained. Our code is available at: https://github.com/MSU-MLSys-Lab/CATE.
****
On Perceptual Lossy Compression: The Cost of Perceptual Reconstruction and An Op timal Training Framework
Zeyu Yan, Fei Wen, Rendong Ying, Chao Ma, Peilin Liu
Lossy compression algorithms are typically designed to achieve the lowest possib le distortion at a given bit rate. However, recent studies show that pursuing hi gh perceptual quality would lead to increase of the lowest achievable distortion (e.g., MSE). This paper provides nontrivial results theoretically revealing tha t, 1) the cost of achieving perfect perception quality is exactly a doubling of

the lowest achievable MSE distortion, 2) an optimal encoder for the “classic” rate-distortion problem is also optimal for the perceptual compression problem, 3) distortion loss is unnecessary for training a perceptual decoder. Further, we propose a novel training framework to achieve the lowest MSE distortion under perfect perception constraint at a given bit rate. This framework uses a GAN with discriminator conditioned on an MSE-optimized encoder, which is superior over the traditional framework using distortion plus adversarial loss. Experiments are provided to verify the theoretical finding and demonstrate the superiority of the proposed training framework.
****
CIFS: Improving Adversarial Robustness of CNNs via Channel-wise Importance-based Feature Selection
Hanshu Yan, Jingfeng Zhang, Gang Niu, Jiashi Feng, Vincent Tan, Masashi Sugiyama
We investigate the adversarial robustness of CNNs from the perspective of channel-wise activations. By comparing normally trained and adversarially trained models, we observe that adversarial training (AT) robustifies CNNs by aligning the channel-wise activations of adversarial data with those of their natural counterparts. However, the channels that are \textit{negatively-relevant} (NR) to predictions are still over-activated when processing adversarial data. Besides, we also observe that AT does not result in similar robustness for all classes. For the robust classes, channels with larger activation magnitudes are usually more \textit{positively-relevant} (PR) to predictions, but this alignment does not hold for the non-robust classes. Given these observations, we hypothesize that suppressing NR channels and aligning PR ones with their relevances further enhances the robustness of CNNs under AT. To examine this hypothesis, we introduce a novel mechanism, \textit{i.e.}, \underline{C}hannel-wise \underline{I}mportance-based \underline{F}eature \underline{S}election (CIFS). The CIFS manipulates channels’ activations of certain layers by generating non-negative multipliers to these channels based on their relevances to predictions. Extensive experiments on benchmark datasets including CIFAR10 and SVHN clearly verify the hypothesis and CIFS’s effectiveness of robustifying CNNs.
****
Exact Gap between Generalization Error and Uniform Convergence in Random Feature Models
Zitong Yang, Yu Bai, Song Mei
Recent work showed that there could be a large gap between the classical uniform convergence bound and the actual test error of zero-training-error predictors (interpolators) such as deep neural networks. To better understand this gap, we study the uniform convergence in the nonlinear random feature model and perform a precise theoretical analysis on how uniform convergence depends on the sample size and the number of parameters. We derive and prove analytical expressions for three quantities in this model: 1) classical uniform convergence over norm balls, 2) uniform convergence over interpolators in the norm ball (recently proposed by \citet{zhou2021uniform}), and 3) the risk of minimum norm interpolator. We show that, in the setting where the classical uniform convergence bound is vacuous (diverges to ∞), uniform convergence over the interpolators still gives a non-trivial bound of the test error of interpolating solutions. We also showcase a different setting where classical uniform convergence bound is non-vacuous, but uniform convergence over interpolators can give an improved sample complexity guarantee. Our result provides a first exact comparison between the test errors and uniform convergence bounds for interpolators beyond simple linear models.
****
Learning Optimal Auctions with Correlated Valuations from Samples
Chunxue Yang, Xiaohui Bei
In single-item auction design, it is well known due to Cremer and McLean that when bidders’ valuations are drawn from a correlated prior distribution, the auctioneer can extract full social surplus as revenue. However, in most real-world applications, the prior is usually unknown and can only be learned from historical data. In this work, we investigate the robustness of the optimal auction with c

orrelated valuations via sample complexity analysis. We prove upper and lower bo
unds on the number of samples from the unknown prior required to learn a (1-epsi
lon)-approximately optimal auction. Our results reinforce the common belief that
 optimal correlated auctions are sensitive to the distribution parameters and ha
rd to learn unless the prior distribution is well-behaved.
****
Tensor Programs IV: Feature Learning in Infinite-Width Neural Networks
Greg Yang, Edward J. Hu
As its width tends to infinity, a deep neural network's behavior under gradient
descent can become simplified and predictable (e.g. given by the Neural Tangent
Kernel (NTK)), if it is parametrized appropriately (e.g. the NTK parametrization
). However, we show that the standard and NTK parametrizations of a neural netwo
rk do not admit infinite-width limits that can *learn* features, which is crucia
l for pretraining and transfer learning such as with BERT. We propose simple mod
ifications to the standard parametrization to allow for feature learning in the
limit. Using the *Tensor Programs* technique, we derive explicit formulas for su
ch limits. On Word2Vec and few-shot learning on Omniglot via MAML, two canonical
 tasks that rely crucially on feature learning, we compute these limits exactly.
 We find that they outperform both NTK baselines and finite-width networks, with
 the latter approaching the infinite-width feature learning performance as width
 increases.
****
LARNet: Lie Algebra Residual Network for Face Recognition
Xiaolong Yang, Xiaohong Jia, Dihong Gong, Dong-Ming Yan, Zhifeng Li, Wei Liu
Face recognition is an important yet challenging problem in computer vision. A m
ajor challenge in practical face recognition applications lies in significant va
riations between profile and frontal faces. Traditional techniques address this
challenge either by synthesizing frontal faces or by pose invariant learning. In
 this paper, we propose a novel method with Lie algebra theory to explore how fa
ce rotation in the 3D space affects the deep feature generation process of convo
lutional neural networks (CNNs). We prove that face rotation in the image space
is equivalent to an additive residual component in the feature space of CNNs, wh
ich is determined solely by the rotation. Based on this theoretical finding, we
further design a Lie Algebraic Residual Network (LARNet) for tackling pose robus
t face recognition. Our LARNet consists of a residual subnet for decoding rotati
on information from input face images, and a gating subnet to learn rotation mag
nitude for controlling the strength of the residual component contributing to th
e feature learning process. Comprehensive experimental evaluations on both front
al-profile face datasets and general face recognition datasets convincingly demo
nstrate that our method consistently outperforms the state-of-the-art ones.
****
BASGD: Buffered Asynchronous SGD for Byzantine Learning
Yi-Rui Yang, Wu-Jun Li
Distributed learning has become a hot research topic due to its wide application
 in cluster-based large-scale learning, federated learning, edge computing and s
o on. Most traditional distributed learning methods typically assume no failure
or attack. However, many unexpected cases, such as communication failure and eve
n malicious attack, may happen in real applications. Hence, Byzantine learning (
BL), which refers to distributed learning with failure or attack, has recently a
ttracted much attention. Most existing BL methods are synchronous, which are imp
ractical in some applications due to heterogeneous or offline workers. In these
cases, asynchronous BL (ABL) is usually preferred. In this paper, we propose a n
ovel method, called buffered asynchronous stochastic gradient descent (BASGD), f
or ABL. To the best of our knowledge, BASGD is the first ABL method that can res
ist malicious attack without storing any instances on server. Compared with thos
e methods which need to store instances on server, BASGD has a wider scope of ap
plication. BASGD is proved to be convergent, and be able to resist failure or at
tack. Empirical results show that BASGD significantly outperforms vanilla asynch
ronous stochastic gradient descent (ASGD) and other ABL baselines when there exi
sts failure or attack on workers.

****
Tensor Programs IIb: Architectural Universality Of Neural Tangent Kernel Trainin
g Dynamics
Greg Yang, Etai Littwin
Yang (2020) recently showed that the Neural Tangent Kernel (NTK) at initializati
on has an infinite-width limit for a large class of architectures including mode
rn staples such as ResNet and Transformers. However, their analysis does not app
ly to training. Here, we show the same neural networks (in the so-called NTK par
ametrization) during training follow a kernel gradient descent dynamics in funct
ion space, where the kernel is the infinite-width NTK. This completes the proof
of the architectural universality of NTK behavior. To achieve this result, we ap
ply the Tensor Programs technique: Write the entire SGD dynamics inside a Tensor
 Program and analyze it via the Master Theorem. To facilitate this proof, we dev
elop a graphical notation for Tensor Programs, which we believe is also an impor
tant contribution toward the pedagogy and exposition of the Tensor Programs tech
nique.
****
Graph Neural Networks Inspired by Classical Iterative Algorithms
Yongyi Yang, Tang Liu, Yangkun Wang, Jinjing Zhou, Quan Gan, Zhewei Wei, Zheng Z
hang, Zengfeng Huang, David Wipf
Despite the recent success of graph neural networks (GNN), common architectures
often exhibit significant limitations, including sensitivity to oversmoothing, l
ong-range dependencies, and spurious edges, e.g., as can occur as a result of gr
aph heterophily or adversarial attacks. To at least partially address these issu
es within a simple transparent framework, we consider a new family of GNN layers
 designed to mimic and integrate the update rules of two classical iterative alg
orithms, namely, proximal gradient descent and iterative reweighted least square
s (IRLS). The former defines an extensible base GNN architecture that is immune
to oversmoothing while nonetheless capturing long-range dependencies by allowing
 arbitrary propagation steps. In contrast, the latter produces a novel attention
 mechanism that is explicitly anchored to an underlying end-to-end energy functi
on, contributing stability with respect to edge uncertainty. When combined we ob
tain an extremely simple yet robust model that we evaluate across disparate scen
arios including standardized benchmarks, adversarially-perturbated graphs, graph
s with heterophily, and graphs involving long-range dependencies. In doing so, w
e compare against SOTA GNN approaches that have been explicitly designed for the
 respective task, achieving competitive or superior node classification accuracy
. Our code is available at https://github.com/FFTYYY/TWIRLS. And for an extended
 version of this work, please see https://arxiv.org/abs/2103.06064.
****
Representation Matters: Offline Pretraining for Sequential Decision Making
Mengjiao Yang, Ofir Nachum
The recent success of supervised learning methods on ever larger offline dataset
s has spurred interest in the reinforcement learning (RL) field to investigate w
hether the same paradigms can be translated to RL algorithms. This research area
, known as offline RL, has largely focused on offline policy optimization, aimin
g to find a return-maximizing policy exclusively from offline data. In this pape
r, we consider a slightly different approach to incorporating offline data into
sequential decision-making. We aim to answer the question, what unsupervised obj
ectives applied to offline datasets are able to learn state representations whic
h elevate performance on downstream tasks, whether those downstream tasks be onl
ine RL, imitation learning from expert demonstrations, or even offline policy op
timization based on the same offline dataset? Through a variety of experiments u
tilizing standard offline RL datasets, we find that the use of pretraining with
unsupervised learning objectives can dramatically improve the performance of pol
icy learning algorithms that otherwise yield mediocre performance on their own.
Extensive ablations further provide insights into what components of these unsup
ervised objectives {–} e.g., reward prediction, continuous or discrete represent
ations, pretraining or finetuning {–} are most important and in which settings.
****

Accelerating Safe Reinforcement Learning with Constraint-mismatched Baseline Policies
Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, Peter J Ramadge
We consider the problem of reinforcement learning when provided with (1) a baseline control policy and (2) a set of constraints that the learner must satisfy. The baseline policy can arise from demonstration data or a teacher agent and may provide useful cues for learning, but it might also be sub-optimal for the task at hand, and is not guaranteed to satisfy the specified constraints, which might encode safety, fairness or other application-specific requirements. In order to safely learn from baseline policies, we propose an iterative policy optimization algorithm that alternates between maximizing expected return on the task, minimizing distance to the baseline policy, and projecting the policy onto the constraint-satisfying set. We analyze our algorithm theoretically and provide a finite-time convergence guarantee. In our experiments on five different control tasks, our algorithm consistently outperforms several state-of-the-art baselines, achieving 10 times fewer constraint violations and 40% higher reward on average.
****
Voice2Series: Reprogramming Acoustic Models for Time Series Classification
Chao-Han Huck Yang, Yun-Yun Tsai, Pin-Yu Chen
Learning to classify time series with limited data is a practical yet challenging problem. Current methods are primarily based on hand-designed feature extraction rules or domain-specific data augmentation. Motivated by the advances in deep speech processing models and the fact that voice data are univariate temporal signals, in this paper we propose Voice2Serie (V2S), a novel end-to-end approach that reprograms acoustic models for time series classification, through input transformation learning and output label mapping. Leveraging the representation learning power of a large-scale pre-trained speech processing model, on 31 different time series tasks we show that V2S outperforms or is on part with state-of-the-art methods on 22 tasks, and improves their average accuracy by 1.72%. We further provide theoretical justification of V2S by proving its population risk is upper bounded by the source risk and a Wasserstein distance accounting for feature alignment via reprogramming. Our results offer new and effective means to time series classification.
****
When All We Need is a Piece of the Pie: A Generic Framework for Optimizing Two-way Partial AUC
Zhiyong Yang, Qianqian Xu, Shilong Bao, Yuan He, Xiaochun Cao, Qingming Huang
The Area Under the ROC Curve (AUC) is a crucial metric for machine learning, which evaluates the average performance over all possible True Positive Rates (TPRs) and False Positive Rates (FPRs). Based on the knowledge that a skillful classifier should simultaneously embrace a high TPR and a low FPR, we turn to study a more general variant called Two-way Partial AUC (TPAUC), where only the region with $\mathsf{TPR} \ge \alpha, \mathsf{FPR} \le \beta$ is included in the area. Moreover, a recent work shows that the TPAUC is essentially inconsistent with the existing Partial AUC metrics where only the FPR range is restricted, opening a new problem to seek solutions to leverage high TPAUC. Motivated by this, we present the first trial in this paper to optimize this new metric. The critical challenge along this course lies in the difficulty of performing gradient-based optimization with end-to-end stochastic training, even with a proper choice of surrogate loss. To address this issue, we propose a generic framework to construct surrogate optimization problems, which supports efficient end-to-end training with deep-learning. Moreover, our theoretical analyses show that: 1) the objective function of the surrogate problems will achieve an upper bound of the original problem under mild conditions, and 2) optimizing the surrogate problems leads to good generalization performance in terms of TPAUC with a high probability. Finally, empirical studies over several benchmark datasets speak to the efficacy of our framework.
****
Rethinking Rotated Object Detection with Gaussian Wasserstein Distance Loss
Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, Qi Tian

Boundary discontinuity and its inconsistency to the final detection metric have been the bottleneck for rotating detection regression loss design. In this paper, we propose a novel regression loss based on Gaussian Wasserstein distance as a fundamental approach to solve the problem. Specifically, the rotated bounding box is converted to a 2-D Gaussian distribution, which enables to approximate the indifferentiable rotational IoU induced loss by the Gaussian Wasserstein distance (GWD) which can be learned efficiently by gradient back-propagation. GWD can still be informative for learning even there is no overlapping between two rotating bounding boxes which is often the case for small object detection. Thanks to its three unique properties, GWD can also elegantly solve the boundary discontinuity and square-like problem regardless how the bounding box is defined. Experiments on five datasets using different detectors show the effectiveness of our approach, and codes are available at https://github.com/yangxue0827/RotationDetection.
****
Delving into Deep Imbalanced Regression
Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, Dina Katabi
Real-world data often exhibit imbalanced distributions, where certain target values have significantly fewer observations. Existing techniques for dealing with imbalanced data focus on targets with categorical indices, i.e., different classes. However, many tasks involve continuous targets, where hard boundaries between classes do not exist. We define Deep Imbalanced Regression (DIR) as learning from such imbalanced data with continuous targets, dealing with potential missing data for certain target values, and generalizing to the entire target range. Motivated by the intrinsic difference between categorical and continuous label space, we propose distribution smoothing for both labels and features, which explicitly acknowledges the effects of nearby targets, and calibrates both label and learned feature distributions. We curate and benchmark large-scale DIR datasets from common real-world tasks in computer vision, natural language processing, and healthcare domains. Extensive experiments verify the superior performance of our strategies. Our work fills the gap in benchmarks and techniques for practical imbalanced regression problems. Code and data are available at: https://github.com/YyzHarry/imbalanced-regression.
****
Backpropagated Neighborhood Aggregation for Accurate Training of Spiking Neural Networks
Yukun Yang, Wenrui Zhang, Peng Li
While Backpropagation (BP) has been applied to spiking neural networks (SNNs) achieving encouraging results, a key challenge involved is to backpropagate a differentiable continuous-valued loss over layers of spiking neurons exhibiting discontinuous all-or-none firing activities. Existing methods deal with this difficulty by introducing compromises that come with their own limitations, leading to potential performance degradation. We propose a novel BP-like method, called neighborhood aggregation (NA), which computes accurate error gradients guiding weight updates that may lead to discontinuous modifications of firing activities. NA achieves this goal by aggregating the error gradient over multiple spike trains in the neighborhood of the present spike train of each neuron. The employed aggregation is based on a generalized finite difference approximation with a proposed distance metric quantifying the similarity between a given pair of spike trains. Our experiments show that the proposed NA algorithm delivers state-of-the-art performance for SNN training on several datasets including CIFAR10.
****
SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks
Lingxiao Yang, Ru-Yuan Zhang, Lida Li, Xiaohua Xie
In this paper, we propose a conceptually simple but very effective attention module for Convolutional Neural Networks (ConvNets). In contrast to existing channel-wise and spatial-wise attention modules, our module instead infers 3-D attention weights for the feature map in a layer without adding parameters to the original networks. Specifically, we base on some well-known neuroscience theories and

propose to optimize an energy function to find the importance of each neuron. We further derive a fast closed-form solution for the energy function, and show that the solution can be implemented in less than ten lines of code. Another advantage of the module is that most of the operators are selected based on the solution to the defined energy function, avoiding too many efforts for structure tuning. Quantitative evaluations on various visual tasks demonstrate that the proposed module is flexible and effective to improve the representation ability of many ConvNets. Our code is available at Pytorch-SimAM.
****
HAWQ-V3: Dyadic Neural Network Quantization
Zhewei Yao, Zhen Dong, Zhangcheng Zheng, Amir Gholami, Jiali Yu, Eric Tan, Leyuan Wang, Qijing Huang, Yida Wang, Michael Mahoney, Kurt Keutzer
Current low-precision quantization algorithms often have the hidden cost of conversion back and forth from floating point to quantized integer values. This hidden cost limits the latency improvement realized by quantizing Neural Networks. To address this, we present HAWQ-V3, a novel mixed-precision integer-only quantization framework. The contributions of HAWQ-V3 are the following: (i) An integer-only inference where the entire computational graph is performed only with integer multiplication, addition, and bit shifting, without any floating point operations or even integer division; (ii) A novel hardware-aware mixed-precision quantization method where the bit-precision is calculated by solving an integer linear programming problem that balances the trade-off between model perturbation and other constraints, e.g., memory footprint and latency; (iii) Direct hardware deployment and open source contribution for 4-bit uniform/mixed-precision quantization in TVM, achieving an average speed up of 1.45x for uniform 4-bit, as compared to uniform 8-bit for ResNet50 on T4 GPUs; and (iv) extensive evaluation of the proposed methods on ResNet18/50 and InceptionV3, for various model compression levels with/without mixed precision. For ResNet50, our INT8 quantization achieves an accuracy of 77.58%, which is 2.68% higher than prior integer-only work, and our mixed-precision INT4/8 quantization can reduce INT8 latency by 23% and still achieve 76.73% accuracy. Our framework and the TVM implementation have been open sourced (HAWQ, 2020).
****
Improving Generalization in Meta-learning via Task Augmentation
Huaxiu Yao, Long-Kai Huang, Linjun Zhang, Ying Wei, Li Tian, James Zou, Junzhou Huang, Zhenhui () Li
Meta-learning has proven to be a powerful paradigm for transferring the knowledge from previous tasks to facilitate the learning of a novel task. Current dominant algorithms train a well-generalized model initialization which is adapted to each task via the support set. The crux lies in optimizing the generalization capability of the initialization, which is measured by the performance of the adapted model on the query set of each task. Unfortunately, this generalization measure, evidenced by empirical results, pushes the initialization to overfit the meta-training tasks, which significantly impairs the generalization and adaptation to novel tasks. To address this issue, we actively augment a meta-training task with “more data” when evaluating the generalization. Concretely, we propose two task augmentation methods, including MetaMix and Channel Shuffle. MetaMix linearly combines features and labels of samples from both the support and query sets. For each class of samples, Channel Shuffle randomly replaces a subset of their channels with the corresponding ones from a different class. Theoretical studies show how task augmentation improves the generalization of meta-learning. Moreover, both MetaMix and Channel Shuffle outperform state-of-the-art results by a large margin across many datasets and are compatible with existing meta-learning algorithms.
****
Deep Learning for Functional Data Analysis with Adaptive Basis Layers
Junwen Yao, Jonas Mueller, Jane-Ling Wang
Despite their widespread success, the application of deep neural networks to functional data remains scarce today. The infinite dimensionality of functional data means standard learning algorithms can be applied only after appropriate dimen

sion reduction, typically achieved via basis expansions. Currently, these bases are chosen a priori without the information for the task at hand and thus may no t be effective for the designated task. We instead propose to adaptively learn t hese bases in an end-to-end fashion. We introduce neural networks that employ a new Basis Layer whose hidden units are each basis functions themselves implement ed as a micro neural network. Our architecture learns to apply parsimonious dime nsion reduction to functional inputs that focuses only on information relevant t o the target rather than irrelevant variation in the input function. Across nume rous classification/regression tasks with functional data, our method empiricall y outperforms other types of neural networks, and we prove that our approach is statistically consistent with low generalization error.
****
Addressing Catastrophic Forgetting in Few-Shot Problems
Pauching Yap, Hippolyt Ritter, David Barber
Neural networks are known to suffer from catastrophic forgetting when trained on sequential datasets. While there have been numerous attempts to solve this prob lem in large-scale supervised classification, little has been done to overcome c atastrophic forgetting in few-shot classification problems. We demonstrate that the popular gradient-based model-agnostic meta-learning algorithm (MAML) indeed suffers from catastrophic forgetting and introduce a Bayesian online meta-learni ng framework that tackles this problem. Our framework utilises Bayesian online l earning and meta-learning along with Laplace approximation and variational infer ence to overcome catastrophic forgetting in few-shot classification problems. Th e experimental evaluations demonstrate that our framework can effectively achiev e this goal in comparison with various baselines. As an additional utility, we a lso demonstrate empirically that our framework is capable of meta-learning on se quentially arriving few-shot tasks from a stationary task distribution.
****
Reinforcement Learning with Prototypical Representations
Denis Yarats, Rob Fergus, Alessandro Lazaric, Lerrel Pinto
Learning effective representations in image-based environments is crucial for sa mple efficient Reinforcement Learning (RL). Unfortunately, in RL, representation learning is confounded with the exploratory experience of the agent – learning a useful representation requires diverse data, while effective exploration is on ly possible with coherent representations. Furthermore, we would like to learn r epresentations that not only generalize across tasks but also accelerate downstr eam exploration for efficient task-specific training. To address these challenge s we propose Proto-RL, a self-supervised framework that ties representation lear ning with exploration through prototypical representations. These prototypes sim ultaneously serve as a summarization of the exploratory experience of an agent a s well as a basis for representing observations. We pre-train these task-agnosti c representations and prototypes on environments without downstream task informa tion. This enables state-of-the-art downstream policy learning on a set of diffi cult continuous control tasks.
****
Elementary superexpressive activations
Dmitry Yarotsky
We call a finite family of activation functions \emph{superexpressive} if any mu ltivariate continuous function can be approximated by a neural network that uses these activations and has a fixed architecture only depending on the number of input variables (i.e., to achieve any accuracy we only need to adjust the weight s, without increasing the number of neurons). Previously, it was known that supe rexpressive activations exist, but their form was quite complex. We give example s of very simple superexpressive families: for example, we prove that the family $\{sin, arcsin\}$ is superexpressive. We also show that most practical activati ons (not involving periodic functions) are not superexpressive.
****
Break-It-Fix-It: Unsupervised Learning for Program Repair
Michihiro Yasunaga, Percy Liang
We consider repair tasks: given a critic (e.g., compiler) that assesses the qual

ity of an input, the goal is to train a fixer that converts a bad example (e.g.,
code with syntax errors) into a good one (e.g., code with no errors). Existing
works create training data consisting of (bad, good) pairs by corrupting good ex
amples using heuristics (e.g., dropping tokens). However, fixers trained on this
synthetically-generated data do not extrapolate well to the real distribution o
f bad inputs. To bridge this gap, we propose a new training approach, Break-It-F
ix-It (BIFI), which has two key ideas: (i) we use the critic to check a fixer’s
output on real bad inputs and add good (fixed) outputs to the training data, and
(ii) we train a breaker to generate realistic bad code from good code. Based on
these ideas, we iteratively update the breaker and the fixer while using them i
n conjunction to generate more paired data. We evaluate BIFI on two code repair
datasets: GitHub-Python, a new dataset we introduce where the goal is to repair
Python code with AST parse errors; and DeepFix, where the goal is to repair C co
de with compiler errors. BIFI outperforms existing methods, obtaining 90.5% repa
ir accuracy on GitHub-Python (+28.5%) and 71.7% on DeepFix (+5.6%). Notably, BIF
I does not require any labeled data; we hope it will be a strong starting point
for unsupervised learning of various repair tasks.
****
Improving Gradient Regularization using Complex-Valued Neural Networks
Eric C Yeats, Yiran Chen, Hai Li
Gradient regularization is a neural network defense technique that requires no p
rior knowledge of an adversarial attack and that brings only limited increase in
training computational complexity. A form of complex-valued neural network (CVN
N) is proposed to improve the performance of gradient regularization on classifi
cation tasks of real-valued input in adversarial settings. The activation deriva
tives of each layer of the CVNN are dependent on the combination of inputs to th
e layer, and locally stable representations can be learned for inputs the networ
k is trained on. Furthermore, the properties of the CVNN parameter derivatives r
esist decrease of performance on the standard objective that is caused by compet
ition with the gradient regularization objective. Experimental results show that
the performance of gradient regularized CVNN surpasses that of real-valued neur
al networks with comparable storage and computational complexity. Moreover, grad
ient regularized complex-valued networks exhibit robust performance approaching
that of real-valued networks trained with multi-step adversarial training.
****
Neighborhood Contrastive Learning Applied to Online Patient Monitoring
Hugo Yèche, Gideon Dresdner, Francesco Locatello, Matthias Hüser, Gunnar Rätsch
Intensive care units (ICU) are increasingly looking towards machine learning for
methods to provide online monitoring of critically ill patients. In machine lea
rning, online monitoring is often formulated as a supervised learning problem. R
ecently, contrastive learning approaches have demonstrated promising improvement
s over competitive supervised benchmarks. These methods rely on well-understood
data augmentation techniques developed for image data which do not apply to onli
ne monitoring. In this work, we overcome this limitation by supplementing time-s
eries data augmentation techniques with a novel contrastive learning objective w
hich we call neighborhood contrastive learning (NCL). Our objective explicitly g
roups together contiguous time segments from each patient while maintaining stat
e-specific information. Our experiments demonstrate a marked improvement over ex
isting work applying contrastive methods to medical time-series.
****
From Local Structures to Size Generalization in Graph Neural Networks
Gilad Yehudai, Ethan Fetaya, Eli Meirom, Gal Chechik, Haggai Maron
Graph neural networks (GNNs) can process graphs of different sizes, but their ab
ility to generalize across sizes, specifically from small to large graphs, is st
ill not well understood. In this paper, we identify an important type of data wh
ere generalization from small to large graphs is challenging: graph distribution
s for which the local structure depends on the graph size. This effect occurs in
multiple important graph learning domains, including social and biological netw
orks. We first prove that when there is a difference between the local structure
s, GNNs are not guaranteed to generalize across sizes: there are "bad" global mi

nima that do well on small graphs but fail on large graphs. We then study the size-generalization problem empirically and demonstrate that when there is a discrepancy in local structure, GNNs tend to converge to non-generalizing solutions. Finally, we suggest two approaches for improving size generalization, motivated by our findings. Notably, we propose a novel Self-Supervised Learning (SSL) task aimed at learning meaningful representations of local structures that appear in large graphs. Our SSL task improves classification accuracy on several popular datasets.
****
Improved OOD Generalization via Adversarial Training and Pretraing
Mingyang Yi, Lu Hou, Jiacheng Sun, Lifeng Shang, Xin Jiang, Qun Liu, Zhiming Ma
Recently, learning a model that generalizes well on out-of-distribution (OOD) data has attracted great attention in the machine learning community. In this paper, after defining OOD generalization by Wasserstein distance, we theoretically justify that a model robust to input perturbation also generalizes well on OOD data. Inspired by previous findings that adversarial training helps improve robustness, we show that models trained by adversarial training have converged excess risk on OOD data. Besides, in the paradigm of pre-training then fine-tuning, we theoretically justify that the input perturbation robust model in the pre-training stage provides an initialization that generalizes well on downstream OOD data. Finally, various experiments conducted on image classification and natural language understanding tasks verify our theoretical findings.
****
Regret and Cumulative Constraint Violation Analysis for Online Convex Optimization with Long Term Constraints
Xinlei Yi, Xiuxian Li, Tao Yang, Lihua Xie, Tianyou Chai, Karl Johansson
This paper considers online convex optimization with long term constraints, where constraints can be violated in intermediate rounds, but need to be satisfied in the long run. The cumulative constraint violation is used as the metric to measure constraint violations, which excludes the situation that strictly feasible constraints can compensate the effects of violated constraints. A novel algorithm is first proposed and it achieves an $\mathcal{O}(T^{\max\{c,1-c\}})$ bound for static regret and an $\mathcal{O}(T^{(1-c)/2})$ bound for cumulative constraint violation, where $c\in(0,1)$ is a user-defined trade-off parameter, and thus has improved performance compared with existing results. Both static regret and cumulative constraint violation bounds are reduced to $\mathcal{O}(\log(T))$ when the loss functions are strongly convex, which also improves existing results. % In order to bound the regret with respect to any comparator sequence, In order to achieve the optimal regret with respect to any comparator sequence, another algorithm is then proposed and it achieves the optimal $\mathcal{O}(\sqrt{T(1+P_T)})$ regret and an $\mathcal{O}(\sqrt{T})$ cumulative constraint violation, where P_T is the path-length of the comparator sequence. Finally, numerical simulations are provided to illustrate the effectiveness of the theoretical results.
****
Continuous-time Model-based Reinforcement Learning
Cagatay Yildiz, Markus Heinonen, Harri Lähdesmäki
Model-based reinforcement learning (MBRL) approaches rely on discrete-time state transition models whereas physical systems and the vast majority of control tasks operate in continuous-time. To avoid time-discretization approximation of the underlying process, we propose a continuous-time MBRL framework based on a novel actor-critic method. Our approach also infers the unknown state evolution differentials with Bayesian neural ordinary differential equations (ODE) to account for epistemic uncertainty. We implement and test our method on a new ODE-RL suite that explicitly solves continuous-time control systems. Our experiments illustrate that the model is robust against irregular and noisy data, and can solve classic control problems in a sample-efficient manner.
****
Distributed Nyström Kernel Learning with Communications
Rong Yin, Weiping Wang, Dan Meng
We study the statistical performance for distributed kernel ridge regression wit

h Nyström (DKRR-NY) and with Nyström and iterative solvers (DKRR-NY-PCG) and successfully derive the optimal learning rates, which can improve the ranges of the number of local processors p to the optimal in existing state-of-art bounds. More precisely, our theoretical analysis show that DKRR-NY and DKRR-NY-PCG achieve the same learning rates as the exact KRR requiring essentially $\mathcal{O}(|D|^{1.5})$ time and $\mathcal{O}(|D|)$ memory with relaxing the restriction on p in expectation, where $|D|$ is the number of data, which exhibits the average effectiveness of multiple trials. Furthermore, for showing the generalization performance in a single trial, we deduce the learning rates for DKRR-NY and DKRR-NY-PCG in probability. Finally, we propose a novel algorithm DKRR-NY-CM based on DKRR-NY, which employs a communication strategy to further improve the learning performance, whose effectiveness of communications is validated in theoretical and experimental analysis.
****

Path Planning using Neural A* Search

Ryo Yonetani, Tatsunori Taniai, Mohammadamin Barekatain, Mai Nishimura, Asako Kanezaki

We present Neural A*, a novel data-driven search method for path planning problems. Despite the recent increasing attention to data-driven path planning, machine learning approaches to search-based planning are still challenging due to the discrete nature of search algorithms. In this work, we reformulate a canonical A* search algorithm to be differentiable and couple it with a convolutional encoder to form an end-to-end trainable neural network planner. Neural A* solves a path planning problem by encoding a problem instance to a guidance map and then performing the differentiable A* search with the guidance map. By learning to match the search results with ground-truth paths provided by experts, Neural A* can produce a path consistent with the ground truth accurately and efficiently. Our extensive experiments confirmed that Neural A* outperformed state-of-the-art data-driven planners in terms of the search optimality and efficiency trade-off. Furthermore, Neural A* successfully predicted realistic human trajectories by directly performing search-based planning on natural image inputs.
****

SinIR: Efficient General Image Manipulation with Single Image Reconstruction

Jihyeong Yoo, Qifeng Chen

We propose SinIR, an efficient reconstruction-based framework trained on a single natural image for general image manipulation, including super-resolution, editing, harmonization, paint-to-image, photo-realistic style transfer, and artistic style transfer. We train our model on a single image with cascaded multi-scale learning, where each network at each scale is responsible for image reconstruction. This reconstruction objective greatly reduces the complexity and running time of training, compared to the GAN objective. However, the reconstruction objective also exacerbates the output quality. Therefore, to solve this problem, we further utilize simple random pixel shuffling, which also gives control over manipulation, inspired by the Denoising Autoencoder. With quantitative evaluation, we show that SinIR has competitive performance on various image manipulation tasks. Moreover, with a much simpler training objective (i.e., reconstruction), SinIR is trained 33.5 times faster than SinGAN (for 500x500 images) that solves similar tasks. Our code is publicly available at github.com/YooJiHyeong/SinIR.
****

Conditional Temporal Neural Processes with Covariance Loss

Boseon Yoo, Jiwoo Lee, Janghoon Ju, Seijun Chung, Soyeon Kim, Jaesik Choi

We introduce a novel loss function, Covariance Loss, which is conceptually equivalent to conditional neural processes and has a form of regularization so that is applicable to many kinds of neural networks. With the proposed loss, mappings from input variables to target variables are highly affected by dependencies of target variables as well as mean activation and mean dependencies of input and target variables. This nature enables the resulting neural networks to become more robust to noisy observations and recapture missing dependencies from prior information. In order to show the validity of the proposed loss, we conduct extensive sets of experiments on real-world datasets with state-of-the-art models and d

iscuss the benefits and drawbacks of the proposed Covariance Loss.
****
Adversarial Purification with Score-based Generative Models
Jongmin Yoon, Sung Ju Hwang, Juho Lee
While adversarial training is considered as a standard defense method against adversarial attacks for image classifiers, adversarial purification, which purifies attacked images into clean images with a standalone purification, model has shown promises as an alternative defense method. Recently, an EBM trained with MCMC has been highlighted as a purification model, where an attacked image is purified by running a long Markov-chain using the gradients of the EBM. Yet, the practicality of the adversarial purification using an EBM remains questionable because the number of MCMC steps required for such purification is too large. In this paper, we propose a novel adversarial purification method based on an EBM trained with DSM. We show that an EBM trained with DSM can quickly purify attacked images within a few steps. We further introduce a simple yet effective randomized purification scheme that injects random noises into images before purification. This process screens the adversarial perturbations imposed on images by the random noises and brings the images to the regime where the EBM can denoise well. We show that our purification method is robust against various attacks and demonstrate its state-of-the-art performances.
****
Federated Continual Learning with Weighted Inter-client Transfer
Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, Sung Ju Hwang
There has been a surge of interest in continual learning and federated learning, both of which are important in deep neural networks in real-world scenarios. Yet little research has been done regarding the scenario where each client learns on a sequence of tasks from a private local data stream. This problem of federated continual learning poses new challenges to continual learning, such as utilizing knowledge from other clients, while preventing interference from irrelevant knowledge. To resolve these issues, we propose a novel federated continual learning framework, Federated Weighted Inter-client Transfer (FedWeIT), which decomposes the network weights into global federated parameters and sparse task-specific parameters, and each client receives selective knowledge from other clients by taking a weighted combination of their task-specific parameters. FedWeIT minimizes interference between incompatible tasks, and also allows positive knowledge transfer across clients during learning. We validate our FedWeIT against existing federated learning and continual learning methods under varying degrees of task similarity across clients, and our model significantly outperforms them with a large reduction in the communication cost.
****
Autoencoding Under Normalization Constraints
Sangwoong Yoon, Yung-Kyun Noh, Frank Park
Likelihood is a standard estimate for outlier detection. The specific role of the normalization constraint is to ensure that the out-of-distribution (OOD) regime has a small likelihood when samples are learned using maximum likelihood. Because autoencoders do not possess such a process of normalization, they often fail to recognize outliers even when they are obviously OOD. We propose the Normalized Autoencoder (NAE), a normalized probabilistic model constructed from an autoencoder. The probability density of NAE is defined using the reconstruction error of an autoencoder, which is differently defined in the conventional energy-based model. In our model, normalization is enforced by suppressing the reconstruction of negative samples, significantly improving the outlier detection performance. Our experimental results confirm the efficacy of NAE, both in detecting outliers and in generating in-distribution samples.
****
Accelerated Algorithms for Smooth Convex-Concave Minimax Problems with O(1/k^2) Rate on Squared Gradient Norm
Taeho Yoon, Ernest K Ryu
In this work, we study the computational complexity of reducing the squared gradient magnitude for smooth minimax optimization problems. First, we present algor

ithms with accelerated $\mathcal{O}(1/k^2)$ last-iterate rates, faster than the existing $\mathcal{O}(1/k)$ or slower rates for extragradient, Popov, and gradie nt descent with anchoring. The acceleration mechanism combines extragradient ste ps with anchoring and is distinct from Nesterov’s acceleration. We then establis h optimality of the $\mathcal{O}(1/k^2)$ rate through a matching lower bound.
****
Lower-Bounded Proper Losses for Weakly Supervised Classification
Shuhei M Yoshida, Takashi Takenouchi, Masashi Sugiyama
This paper discusses the problem of weakly supervised classification, in which i nstances are given weak labels that are produced by some label-corruption proces s. The goal is to derive conditions under which loss functions for weak-label le arning are proper and lower-bounded—two essential requirements for the losses us ed in class-probability estimation. To this end, we derive a representation theo rem for proper losses in supervised learning, which dualizes the Savage represen tation. We use this theorem to characterize proper weak-label losses and find a condition for them to be lower-bounded. From these theoretical findings, we deri ve a novel regularization scheme called generalized logit squeezing, which makes any proper weak-label loss bounded from below, without losing properness. Furth ermore, we experimentally demonstrate the effectiveness of our proposed approach , as compared to improper or unbounded losses. The results highlight the importa nce of properness and lower-boundedness.
****
Graph Contrastive Learning Automated
Yuning You, Tianlong Chen, Yang Shen, Zhangyang Wang
Self-supervised learning on graph-structured data has drawn recent interest for learning generalizable, transferable and robust representations from unlabeled g raphs. Among many, graph contrastive learning (GraphCL) has emerged with promisi ng representation learning performance. Unfortunately, unlike its counterpart on image data, the effectiveness of GraphCL hinges on ad-hoc data augmentations, w hich have to be manually picked per dataset, by either rules of thumb or trial-a nd-errors, owing to the diverse nature of graph data. That significantly limits the more general applicability of GraphCL. Aiming to fill in this crucial gap, t his paper proposes a unified bi-level optimization framework to automatically, a daptively and dynamically select data augmentations when performing GraphCL on s pecific graph data. The general framework, dubbed JOint Augmentation Optimizatio n (JOAO), is instantiated as min-max optimization. The selections of augmentatio ns made by JOAO are shown to be in general aligned with previous "best practices " observed from handcrafted tuning: yet now being automated, more flexible and v ersatile. Moreover, we propose a new augmentation-aware projection head mechanis m, which will route output features through different projection heads correspon ding to different augmentations chosen at each training step. Extensive experime nts demonstrate that JOAO performs on par with or sometimes better than the stat e-of-the-art competitors including GraphCL, on multiple graph datasets of variou s scales and types, yet without resorting to any laborious dataset-specific tuni ng on augmentation selection. We release the code at https://github.com/Shen-Lab /GraphCL_Automated.
****
LogME: Practical Assessment of Pre-trained Models for Transfer Learning
Kaichao You, Yong Liu, Jianmin Wang, Mingsheng Long
This paper studies task adaptive pre-trained model selection, an underexplored p roblem of assessing pre-trained models for the target task and select best ones from the model zoo \emph{without fine-tuning}. A few pilot works addressed the p roblem in transferring supervised pre-trained models to classification tasks, bu t they cannot handle emerging unsupervised pre-trained models or regression task s. In pursuit of a practical assessment method, we propose to estimate the maxim um value of label evidence given features extracted by pre-trained models. Unlik e the maximum likelihood, the maximum evidence is \emph{immune to over-fitting}, while its expensive computation can be dramatically reduced by our carefully de signed algorithm. The Logarithm of Maximum Evidence (LogME) can be used to asses s pre-trained models for transfer learning: a pre-trained model with a high LogM

E value is likely to have good transfer performance. LogME is \emph{fast, accura
te, and general}, characterizing itself as the first practical method for assess
ing pre-trained models. Compared with brute-force fine-tuning, LogME brings at m
ost $3000\times$ speedup in wall-clock time and requires only $1%$ memory footpr
int. It outperforms prior methods by a large margin in their setting and is appl
icable to new settings. It is general enough for diverse pre-trained models (sup
ervised pre-trained and unsupervised pre-trained), downstream tasks (classificat
ion and regression), and modalities (vision and language). Code is available at
this repository: \href{https://github.com/thuml/LogME}{https://github.com/thuml/
LogME}.
****
Exponentially Many Local Minima in Quantum Neural Networks
Xuchen You, Xiaodi Wu
Quantum Neural Networks (QNNs), or the so-called variational quantum circuits, a
re important quantum applications both because of their similar promises as clas
sical neural networks and because of the feasibility of their implementation on
near-term intermediate-size noisy quantum machines (NISQ). However, the training
 task of QNNs is challenging and much less understood. We conduct a quantitative
 investigation on the landscape of loss functions of QNNs and identify a class o
f simple yet extremely hard QNN instances for training. Specifically, we show fo
r typical under-parameterized QNNs, there exists a dataset that induces a loss f
unction with the number of spurious local minima depending exponentially on the
number of parameters. Moreover, we show the optimality of our construction by pr
oviding an almost matching upper bound on such dependence. While local minima in
 classical neural networks are due to non-linear activations, in quantum neural
networks local minima appear as a result of the quantum interference phenomenon.
 Finally, we empirically confirm that our constructions can indeed be hard insta
nces in practice with typical gradient-based optimizers, which demonstrates the
practical value of our findings.
****
DAGs with No Curl: An Efficient DAG Structure Learning Approach
Yue Yu, Tian Gao, Naiyu Yin, Qiang Ji
Recently directed acyclic graph (DAG) structure learning is formulated as a cons
trained continuous optimization problem with continuous acyclicity constraints a
nd was solved iteratively through subproblem optimization. To further improve ef
ficiency, we propose a novel learning framework to model and learn the weighted
adjacency matrices in the DAG space directly. Specifically, we first show that t
he set of weighted adjacency matrices of DAGs are equivalent to the set of weigh
ted gradients of graph potential functions, and one may perform structure learni
ng by searching in this equivalent set of DAGs. To instantiate this idea, we pro
pose a new algorithm, DAG-NoCurl, which solves the optimization problem efficien
tly with a two-step procedure: $1)$ first we find an initial non-acyclic solutio
n to the optimization problem, and $2)$ then we employ the Hodge decomposition o
f graphs and learn an acyclic graph by projecting the non-acyclic graph to the g
radient of a potential function. Experimental studies on benchmark datasets demo
nstrate that our method provides comparable accuracy but better efficiency than
baseline DAG structure learning methods on both linear and generalized structura
l equation models, often by more than one order of magnitude.
****
Provably Efficient Algorithms for Multi-Objective Competitive RL
Tiancheng Yu, Yi Tian, Jingzhao Zhang, Suvrit Sra
We study multi-objective reinforcement learning (RL) where an agent’s reward is
represented as a vector. In settings where an agent competes against opponents,
its performance is measured by the distance of its average return vector to a ta
rget set. We develop statistically and computationally efficient algorithms to a
pproach the associated target set. Our results extend Blackwell’s approachabilit
y theorem \citep{blackwell1956analog} to tabular RL, where strategic exploration
 becomes essential. The algorithms presented are adaptive; their guarantees hold
 even without Blackwell’s approachability condition. If the opponents use fixed
policies, we give an improved rate of approaching the target set while also tack

ling the more ambitious goal of simultaneously minimizing a scalar cost function . We discuss our analysis for this special case by relating our results to previous works on constrained RL. To our knowledge, this work provides the first provably efficient algorithms for vector-valued Markov games and our theoretical guarantees are near-optimal.
****

Whittle Networks: A Deep Likelihood Model for Time Series
Zhongjie Yu, Fabrizio G Ventola, Kristian Kersting
While probabilistic circuits have been extensively explored for tabular data, less attention has been paid to time series. Here, the goal is to estimate joint densities among the entire time series and, in turn, determining, for instance, conditional independence relations between them. To this end, we propose the first probabilistic circuits (PCs) approach for modeling the joint distribution of multivariate time series, called Whittle sum-product networks (WSPNs). WSPNs leverage the Whittle approximation, casting the likelihood in the frequency domain, and place a complex-valued sum-product network, the most prominent PC, over the frequencies. The conditional independence relations among the time series can then be determined efficiently in the spectral domain. Moreover, WSPNs can naturally be placed into the deep neural learning stack for time series, resulting in Whittle Networks, opening the likelihood toolbox for training deep neural models and inspecting their behaviour. Our experiments show that Whittle Networks can indeed capture complex dependencies between time series and provide a useful measure of uncertainty for neural networks.
****

Deep Latent Graph Matching
Tianshu Yu, Runzhong Wang, Junchi Yan, Baoxin Li
Deep learning for graph matching (GM) has emerged as an important research topic due to its superior performance over traditional methods and insights it provides for solving other combinatorial problems on graph. While recent deep methods for GM extensively investigated effective node/edge feature learning or downstream GM solvers given such learned features, there is little existing work questioning if the fixed connectivity/topology typically constructed using heuristics (e.g., Delaunay or k-nearest) is indeed suitable for GM. From a learning perspective, we argue that the fixed topology may restrict the model capacity and thus potentially hinder the performance. To address this, we propose to learn the (distribution of) latent topology, which can better support the downstream GM task. We devise two latent graph generation procedures, one deterministic and one generative. Particularly, the generative procedure emphasizes the across-graph consistency and thus can be viewed as a matching-guided co-generative model. Our methods deliver superior performance over previous state-of-the-arts on public benchmarks, hence supporting our hypothesis.
****

Learning Generalized Intersection Over Union for Dense Pixelwise Prediction
Jiaqian Yu, Jingtao Xu, Yiwei Chen, Weiming Li, Qiang Wang, Byungin Yoo, Jae-Joon Han
Intersection over union (IoU) score, also named Jaccard Index, is one of the most fundamental evaluation methods in machine learning. The original IoU computation cannot provide non-zero gradients and thus cannot be directly optimized by nowadays deep learning methods. Several recent works generalized IoU for bounding box regression, but they are not straightforward to adapt for pixelwise prediction. In particular, the original IoU fails to provide effective gradients for the non-overlapping and location-deviation cases, which results in performance plateau. In this paper, we propose PixIoU, a generalized IoU for pixelwise prediction that is sensitive to the distance for non-overlapping cases and the locations in prediction. We provide proofs that PixIoU holds many nice properties as the original IoU. To optimize the PixIoU, we also propose a loss function that is proved to be submodular, hence we can apply the Lovász functions, the efficient surrogates for submodular functions for learning this loss. Experimental results show consistent performance improvements by learning PixIoU over the original IoU for several different pixelwise prediction tasks on Pascal VOC, VOT-2020 and Cit

yscapes.
****
Large Scale Private Learning via Low-rank Reparametrization
Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, Tie-Yan Liu
We propose a reparametrization scheme to address the challenges of applying differentially private SGD on large neural networks, which are 1) the huge memory cost of storing individual gradients, 2) the added noise suffering notorious dimensional dependence. Specifically, we reparametrize each weight matrix with two \emph{gradient-carrier} matrices of small dimension and a \emph{residual weight} matrix. We argue that such reparametrization keeps the forward/backward process unchanged while enabling us to compute the projected gradient without computing the gradient itself. To learn with differential privacy, we design \emph{reparametrized gradient perturbation (RGP)} that perturbs the gradients on gradient-carrier matrices and reconstructs an update for the original weight from the noisy gradients. Importantly, we use historical updates to find the gradient-carrier matrices, whose optimality is rigorously justified under linear regression and empirically verified with deep learning tasks. RGP significantly reduces the memory cost and improves the utility. For example, we are the first able to apply differential privacy on the BERT model and achieve an average accuracy of $83.9%$ on four downstream tasks with $\epsilon=8$, which is within $5%$ loss compared to the non-private baseline but enjoys much lower privacy leakage risk.
****
Federated Deep AUC Maximization for Hetergeneous Data with a Constant Communication Complexity
Zhuoning Yuan, Zhishuai Guo, Yi Xu, Yiming Ying, Tianbao Yang
Deep AUC (area under the ROC curve) Maximization (DAM) has attracted much attention recently due to its great potential for imbalanced data classification. However, the research on Federated Deep AUC Maximization (FDAM) is still limited. Compared with standard federated learning (FL) approaches that focus on decomposable minimization objectives, FDAM is more complicated due to its minimization objective is non-decomposable over individual examples. In this paper, we propose improved FDAM algorithms for heterogeneous data by solving the popular non-convex strongly-concave min-max formulation of DAM in a distributed fashion, which can also be applied to a class of non-convex strongly-concave min-max problems. A striking result of this paper is that the communication complexity of the proposed algorithm is a constant independent of the number of machines and also independent of the accuracy level, which improves an existing result by orders of magnitude. The experiments have demonstrated the effectiveness of our FDAM algorithm on benchmark datasets, and on medical chest X-ray images from different organizations. Our experiment shows that the performance of FDAM using data from multiple hospitals can improve the AUC score on testing data from a single hospital for detecting life-threatening diseases based on chest radiographs.
****
Neural Tangent Generalization Attacks
Chia-Hung Yuan, Shan-Hung Wu
The remarkable performance achieved by Deep Neural Networks (DNNs) in many applications is followed by the rising concern about data privacy and security. Since DNNs usually require large datasets to train, many practitioners scrape data from external sources such as the Internet. However, an external data owner may not be willing to let this happen, causing legal or ethical issues. In this paper, we study the generalization attacks against DNNs, where an attacker aims to slightly modify training data in order to spoil the training process such that a trained network lacks generalizability. These attacks can be performed by data owners and protect data from unexpected use. However, there is currently no efficient generalization attack against DNNs due to the complexity of a bilevel optimization involved. We propose the Neural Tangent Generalization Attack (NTGA) that, to the best of our knowledge, is the first work enabling clean-label, black-box generalization attack against DNNs. We conduct extensive experiments, and the empirical results demonstrate the effectiveness of NTGA. Our code and perturbed datasets are available at: https://github.com/lionelmessi6410/ntga.

****
On Explainability of Graph Neural Networks via Subgraph Explorations
Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, Shuiwang Ji
We consider the problem of explaining the predictions of graph neural networks (GNNs), which otherwise are considered as black boxes. Existing methods invariably focus on explaining the importance of graph nodes or edges but ignore the substructures of graphs, which are more intuitive and human-intelligible. In this work, we propose a novel method, known as SubgraphX, to explain GNNs by identifying important subgraphs. Given a trained GNN model and an input graph, our SubgraphX explains its predictions by efficiently exploring different subgraphs with Monte Carlo tree search. To make the tree search more effective, we propose to use Shapley values as a measure of subgraph importance, which can also capture the interactions among different subgraphs. To expedite computations, we propose efficient approximation schemes to compute Shapley values for graph data. Our work represents the first attempt to explain GNNs via identifying subgraphs explicitly and directly. Experimental results show that our SubgraphX achieves significantly improved explanations, while keeping computations at a reasonable level.
****
Federated Composite Optimization
Honglin Yuan, Manzil Zaheer, Sashank Reddi
Federated Learning (FL) is a distributed learning paradigm that scales on-device learning collaboratively and privately. Standard FL algorithms such as FEDAVG are primarily geared towards smooth unconstrained settings. In this paper, we study the Federated Composite Optimization (FCO) problem, in which the loss function contains a non-smooth regularizer. Such problems arise naturally in FL applications that involve sparsity, low-rank, monotonicity, or more general constraints. We first show that straightforward extensions of primal algorithms such as FedAvg are not well-suited for FCO since they suffer from the "curse of primal averaging," resulting in poor convergence. As a solution, we propose a new primal-dual algorithm, Federated Dual Averaging (FedDualAvg), which by employing a novel server dual averaging procedure circumvents the curse of primal averaging. Our theoretical analysis and empirical experiments demonstrate that FedDualAvg outperforms the other baselines.
****
Three Operator Splitting with a Nonconvex Loss Function
Alp Yurtsever, Varun Mangalick, Suvrit Sra
We consider the problem of minimizing the sum of three functions, one of which is nonconvex but differentiable, and the other two are convex but possibly nondifferentiable. We investigate the Three Operator Splitting method (TOS) of Davis & Yin (2017) with an aim to extend its theoretical guarantees for this nonconvex problem template. In particular, we prove convergence of TOS with nonasymptotic bounds on its nonstationarity and infeasibility errors. In contrast with the existing work on nonconvex TOS, our guarantees do not require additional smoothness assumptions on the terms comprising the objective; hence they cover instances of particular interest where the nondifferentiable terms are indicator functions. We also extend our results to a stochastic setting where we have access only to an unbiased estimator of the gradient. Finally, we illustrate the effectiveness of the proposed method through numerical experiments on quadratic assignment problems.
****
Grey-box Extraction of Natural Language Models
Santiago Zanella-Beguelin, Shruti Tople, Andrew Paverd, Boris Köpf
Model extraction attacks attempt to replicate a target machine learning model by querying its inference API. State-of-the-art attacks are learning-based and construct replicas by supervised training on the target model’s predictions, but an emerging class of attacks exploit algebraic properties to obtain high-fidelity replicas using orders of magnitude fewer queries. So far, these algebraic attacks have been limited to neural networks with few hidden layers and ReLU activations. In this paper we present algebraic and hybrid algebraic/learning-based attacks on large-scale natural language models. We consider a grey-box setting, targe

ting models with a pre-trained (public) encoder followed by a single (private) classification layer. Our key findings are that (i) with a frozen encoder, high-fidelity extraction is possible with a small number of in-distribution queries, making extraction attacks indistinguishable from legitimate use; (ii) when the encoder is fine-tuned, a hybrid learning-based/algebraic attack improves over the learning-based state-of-the-art without requiring additional queries.
****
Exponential Lower Bounds for Batch Reinforcement Learning: Batch RL can be Exponentially Harder than Online RL
Andrea Zanette
Several practical applications of reinforcement learning involve an agent learning from past data without the possibility of further exploration. Often these applications require us to 1) identify a near optimal policy or to 2) estimate the value of a target policy. For both tasks we derive exponential information-theoretic lower bounds in discounted infinite horizon MDPs with a linear function representation for the action value function even if 1) realizability holds, 2) the batch algorithm observes the exact reward and transition functions, and 3) the batch algorithm is given the best a priori data distribution for the problem class. Our work introduces a new ‘oracle + batch algorithm’ framework to prove lower bounds that hold for every distribution. The work shows an exponential separation between batch and online reinforcement learning.
****
Learning Binary Decision Trees by Argmin Differentiation
Valentina Zantedeschi, Matt Kusner, Vlad Niculae
We address the problem of learning binary decision trees that partition data for some downstream task. We propose to learn discrete parameters (i.e., for tree traversals and node pruning) and continuous parameters (i.e., for tree split functions and prediction functions) simultaneously using argmin differentiation. We do so by sparsely relaxing a mixed-integer program for the discrete parameters, to allow gradients to pass through the program to continuous parameters. We derive customized algorithms to efficiently compute the forward and backward passes. This means that our tree learning procedure can be used as an (implicit) layer in arbitrary deep networks, and can be optimized with arbitrary loss functions. We demonstrate that our approach produces binary trees that are competitive with existing single tree and ensemble approaches, in both supervised and unsupervised settings. Further, apart from greedy approaches (which do not have competitive accuracies), our method is faster to train than all other tree-learning baselines we compare with.
****
Barlow Twins: Self-Supervised Learning via Redundancy Reduction
Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, Stephane Deny
Self-supervised learning (SSL) is rapidly closing the gap with supervised methods on large computer vision benchmarks. A successful approach to SSL is to learn embeddings which are invariant to distortions of the input sample. However, a recurring issue with this approach is the existence of trivial constant solutions. Most current methods avoid such solutions by careful implementation details. We propose an objective function that naturally avoids collapse by measuring the cross-correlation matrix between the outputs of two identical networks fed with distorted versions of a sample, and making it as close to the identity matrix as possible. This causes the embedding vectors of distorted versions of a sample to be similar, while minimizing the redundancy between the components of these vectors. The method is called Barlow Twins, owing to neuroscientist H. Barlow’s redundancy-reduction principle applied to a pair of identical networks. Barlow Twins does not require large batches nor asymmetry between the network twins such as a predictor network, gradient stopping, or a moving average on the weight updates. Intriguingly it benefits from very high-dimensional output vectors. Barlow Twins outperforms previous methods on ImageNet for semi-supervised classification in the low-data regime, and is on par with current state of the art for ImageNet classification with a linear classifier head, and for transfer tasks of classification and object detection.

****
You Only Sample (Almost) Once: Linear Cost Self-Attention Via Bernoulli Sampling
Zhanpeng Zeng, Yunyang Xiong, Sathya Ravi, Shailesh Acharya, Glenn M Fung, Vikas Singh
Transformer-based models are widely used in natural language processing (NLP). Central to the transformer model is the self-attention mechanism, which captures the interactions of token pairs in the input sequences and depends quadratically on the sequence length. Training such models on longer sequences is expensive. In this paper, we show that a Bernoulli sampling attention mechanism based on Locality Sensitive Hashing (LSH), decreases the quadratic complexity of such models to linear. We bypass the quadratic cost by considering self-attention as a sum of individual tokens associated with Bernoulli random variables that can, in principle, be sampled at once by a single hash (although in practice, this number may be a small constant). This leads to an efficient sampling scheme to estimate self-attention which relies on specific modifications of LSH (to enable deployment on GPU architectures). We evaluate our algorithm on the GLUE benchmark with standard 512 sequence length where we see favorable performance relative to a standard pretrained Transformer. On the Long Range Arena (LRA) benchmark, for evaluating performance on long sequences, our method achieves results consistent with softmax self-attention but with sizable speed-ups and memory savings and often outperforms other efficient self-attention methods. Our code is available at https://github.com/mlpen/YOSO.
****
DouZero: Mastering DouDizhu with Self-Play Deep Reinforcement Learning
Daochen Zha, Jingru Xie, Wenye Ma, Sheng Zhang, Xiangru Lian, Xia Hu, Ji Liu
Games are abstractions of the real world, where artificial agents learn to compete and cooperate with other agents. While significant achievements have been made in various perfect- and imperfect-information games, DouDizhu (a.k.a. Fighting the Landlord), a three-player card game, is still unsolved. DouDizhu is a very challenging domain with competition, collaboration, imperfect information, large state space, and particularly a massive set of possible actions where the legal actions vary significantly from turn to turn. Unfortunately, modern reinforcement learning algorithms mainly focus on simple and small action spaces, and not surprisingly, are shown not to make satisfactory progress in DouDizhu. In this work, we propose a conceptually simple yet effective DouDizhu AI system, namely DouZero, which enhances traditional Monte-Carlo methods with deep neural networks, action encoding, and parallel actors. Starting from scratch in a single server with four GPUs, DouZero outperformed all the existing DouDizhu AI programs in days of training and was ranked the first in the Botzone leaderboard among 344 AI agents. Through building DouZero, we show that classic Monte-Carlo methods can be made to deliver strong results in a hard domain with a complex action space. The code and an online demo are released at https://github.com/kwai/DouZero with the hope that this insight could motivate future work.
****
DORO: Distributional and Outlier Robust Optimization
Runtian Zhai, Chen Dan, Zico Kolter, Pradeep Ravikumar
Many machine learning tasks involve subpopulation shift where the testing data distribution is a subpopulation of the training distribution. For such settings, a line of recent work has proposed the use of a variant of empirical risk minimization(ERM) known as distributionally robust optimization (DRO). In this work, we apply DRO to real, large-scale tasks with subpopulation shift, and observe that DRO performs relatively poorly, and moreover has severe instability. We identify one direct cause of this phenomenon: sensitivity of DRO to outliers in the datasets. To resolve this issue, we propose the framework of DORO, for Distributional and Outlier Robust Optimization. At the core of this approach is a refined risk function which prevents DRO from overfitting to potential outliers. We instantiate DORO for the Cressie-Read family of Rényi divergence, and delve into two specific instances of this family: CVaR and χ^2-DRO. We theoretically prove the effectiveness of the proposed method, and empirically show that DORO improves the performance and stability of DRO with experiments on large modern dataset

s, thereby positively addressing the open question raised by Hashimoto et al., 2018. Codes are available at https://github.com/RuntianZ/doro.
****
Can Subnetwork Structure Be the Key to Out-of-Distribution Generalization?
Dinghuai Zhang, Kartik Ahuja, Yilun Xu, Yisen Wang, Aaron Courville
Can models with particular structure avoid being biased towards spurious correlation in out-of-distribution (OOD) generalization? Peters et al. (2016) provides a positive answer for linear cases. In this paper, we use a functional modular probing method to analyze deep model structures under OOD setting. We demonstrate that even in biased models (which focus on spurious correlation) there still exist unbiased functional subnetworks. Furthermore, we articulate and confirm the functional lottery ticket hypothesis: the full network contains a subnetwork with proper structure that can achieve better OOD performance. We then propose Modular Risk Minimization to solve the subnetwork selection problem. Our algorithm learns the functional structure from a given dataset, and can be combined with any other OOD regularization methods. Experiments on various OOD generalization tasks corroborate the effectiveness of our method.
****
Towards Certifying L-infinity Robustness using Neural Networks with L-inf-dist Neurons
Bohang Zhang, Tianle Cai, Zhou Lu, Di He, Liwei Wang
It is well-known that standard neural networks, even with a high classification accuracy, are vulnerable to small ℓ_∞-norm bounded adversarial perturbations. Although many attempts have been made, most previous works either can only provide empirical verification of the defense to a particular attack method, or can only develop a certified guarantee of the model robustness in limited scenarios. In this paper, we seek for a new approach to develop a theoretically principled neural network that inherently resists ℓ_∞ perturbations. In particular, we design a novel neuron that uses ℓ_∞-distance as its basic operation (which we call ℓ_∞-dist neuron), and show that any neural network constructed with ℓ_∞-dist neurons (called ℓ_{∞}-dist net) is naturally a 1-Lipschitz function with respect to ℓ_∞-norm. This directly provides a rigorous guarantee of the certified robustness based on the margin of prediction outputs. We then prove that such networks have enough expressive power to approximate any 1-Lipschitz function with robust generalization guarantee. We further provide a holistic training strategy that can greatly alleviate optimization difficulties. Experimental results show that using ℓ_{∞}-dist nets as basic building blocks, we consistently achieve state-of-the-art performance on commonly used datasets: 93.09% certified accuracy on MNIST ($\epsilon=0.3$), 35.42% on CIFAR-10 ($\epsilon=8/255$) and 16.31% on TinyImageNet ($\epsilon=1/255$).
****
Efficient Lottery Ticket Finding: Less Data is More
Zhenyu Zhang, Xuxi Chen, Tianlong Chen, Zhangyang Wang
The lottery ticket hypothesis (LTH) reveals the existence of winning tickets (sparse but critical subnetworks) for dense networks, that can be trained in isolation from random initialization to match the latter’s accuracies. However, finding winning tickets requires burdensome computations in the train-prune-retrain process, especially on large-scale datasets (e.g., ImageNet), restricting their practical benefits. This paper explores a new perspective on finding lottery tickets more efficiently, by doing so only with a specially selected subset of data, called Pruning-Aware Critical set (PrAC set), rather than using the full training set. The concept of PrAC set was inspired by the recent observation, that deep networks have samples that are either hard to memorize during training, or easy to forget during pruning. A PrAC set is thus hypothesized to capture those most challenging and informative examples for the dense model. We observe that a high-quality winning ticket can be found with training and pruning the dense network on the very compact PrAC set, which can substantially save training iterations for the ticket finding process. Extensive experiments validate our proposal across diverse datasets and network architectures. Specifically, on CIFAR-10, CIFAR

-100, and Tiny ImageNet, we locate effective PrAC sets at 35.32% 78.19% of their training set sizes. On top of them, we can obtain the same competitive winning tickets for the corresponding dense networks, yet saving up to 82.85% 92.77%, 63.54% 74.92%, and 76.14% 86.56% training iterations, respectively. Crucially, we show that a PrAC set found is reusable across different network architectures, which can amortize the extra cost of finding PrAC sets, yielding a practical regime for efficient lottery ticket finding.
****
Robust Policy Gradient against Strong Data Corruption
Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, Wen Sun
We study the problem of robust reinforcement learning under adversarial corruption on both rewards and transitions. Our attack model assumes an \textit{adaptive} adversary who can arbitrarily corrupt the reward and transition at every step within an episode, for at most ϵ-fraction of the learning episodes. Our attack model is strictly stronger than those considered in prior works. Our first result shows that no algorithm can find a better than $O(\epsilon)$-optimal policy under our attack model. Next, we show that surprisingly the natural policy gradient (NPG) method retains a natural robustness property if the reward corruption is bounded, and can find an $O(\sqrt{\epsilon})$-optimal policy. Consequently, we develop a Filtered Policy Gradient (FPG) algorithm that can tolerate even unbounded reward corruption and can find an $O(\epsilon^{1/4})$-optimal policy. We emphasize that FPG is the first that can achieve a meaningful learning guarantee when a constant fraction of episodes are corrupted. Complimentary to the theoretical results, we show that a neural implementation of FPG achieves strong robust learning performance on the MuJoCo continuous control benchmarks.
****
Near Optimal Reward-Free Reinforcement Learning
Zihan Zhang, Simon Du, Xiangyang Ji
We study the reward-free reinforcement learning framework, which is particularly suitable for batch reinforcement learning and scenarios where one needs policies for multiple reward functions. This framework has two phases: in the exploration phase, the agent collects trajectories by interacting with the environment without using any reward signal; in the planning phase, the agent needs to return a near-optimal policy for arbitrary reward functions. %This framework is suitable for batch RL setting and the setting where there are multiple reward functions of interes We give a new efficient algorithm, \textbf{S}taged \textbf{S}ampling + \textbf{T}runcated \textbf{P}lanning (\algoname), which interacts with the environment at most $O\left(\frac{S^2A}{\epsilon^2}\poly\log\left(\frac{SAH}{\epsilon}\right) \right)$ episodes in the exploration phase, and guarantees to output a near-optimal policy for arbitrary reward functions in the planning phase, where S is the size of state space, A is the size of action space, H is the planning horizon, and ϵ is the target accuracy relative to the total reward. Notably, our sample complexity scales only \emph{logarithmically} with H, in contrast to all existing results which scale \emph{polynomially} with H. Furthermore, this bound matches the minimax lower bound $\Omega\left(\frac{S^2A}{\epsilon^2}\right)$ up to logarithmic factors. Our results rely on three new techniques : 1) A new sufficient condition for the dataset to plan for an ϵ-suboptimal policy % for any totally bounded reward function ; 2) A new way to plan efficiently under the proposed condition using soft-truncated planning; 3) Constructing extended MDP to maximize the truncated accumulative rewards efficiently.
****
Bayesian Attention Belief Networks
Shujian Zhang, Xinjie Fan, Bo Chen, Mingyuan Zhou
Attention-based neural networks have achieved state-of-the-art results on a wide range of tasks. Most such models use deterministic attention while stochastic attention is less explored due to the optimization difficulties or complicated model design. This paper introduces Bayesian attention belief networks, which construct a decoder network by modeling unnormalized attention weights with a hierarchy of gamma distributions, and an encoder network by stacking Weibull distribut

ions with a deterministic-upward-stochastic-downward structure to approximate th
e posterior. The resulting auto-encoding networks can be optimized in a differen
tiable way with a variational lower bound. It is simple to convert any models wi
th deterministic attention, including pretrained ones, to the proposed Bayesian
attention belief networks. On a variety of language understanding tasks, we show
 that our method outperforms deterministic attention and state-of-the-art stocha
stic attention in accuracy, uncertainty estimation, generalization across domain
s, and robustness to adversarial attacks. We further demonstrate the general app
licability of our method on neural machine translation and visual question answe
ring, showing great potential of incorporating our method into various attention
-related tasks.
****
Understanding Failures in Out-of-Distribution Detection with Deep Generative Mod
els
Lily Zhang, Mark Goldstein, Rajesh Ranganath
Deep generative models (DGMs) seem a natural fit for detecting out-of-distributi
on (OOD) inputs, but such models have been shown to assign higher probabilities
or densities to OOD images than images from the training distribution. In this w
ork, we explain why this behavior should be attributed to model misestimation. W
e first prove that no method can guarantee performance beyond random chance with
out assumptions on which out-distributions are relevant. We then interrogate the
 typical set hypothesis, the claim that relevant out-distributions can lie in hi
gh likelihood regions of the data distribution, and that OOD detection should be
 defined based on the data distribution’s typical set. We highlight the conseque
nces implied by assuming support overlap between in- and out-distributions, as w
ell as the arbitrariness of the typical set for OOD detection. Our results sugge
st that estimation error is a more plausible explanation than the misalignment b
etween likelihood-based OOD detection and out-distributions of interest, and we
illustrate how even minimal estimation error can lead to OOD detection failures,
 yielding implications for future work in deep generative modeling and OOD detec
tion.
****
Poolingformer: Long Document Modeling with Pooling Attention
Hang Zhang, Yeyun Gong, Yelong Shen, Weisheng Li, Jiancheng Lv, Nan Duan, Weizhu
 Chen
In this paper, we introduce a two-level attention schema, Poolingformer, for lon
g document modeling. Its first level uses a smaller sliding window pattern to ag
gregate information from neighbors. Its second level employs a larger window to
increase receptive fields with pooling attention to reduce both computational co
st and memory consumption. We first evaluate Poolingformer on two long sequence
QA tasks: the monolingual NQ and the multilingual TyDi QA. Experimental results
show that Poolingformer sits atop three official leaderboards measured by F1, ou
tperforming previous state-of-the-art models by 1.9 points (79.8 vs. 77.9) on NQ
 long answer, 1.9 points (79.5 vs. 77.6) on TyDi QA passage answer, and 1.6 poin
ts (67.6 vs. 66.0) on TyDi QA minimal answer. We further evaluate Poolingformer
on a long sequence summarization task. Experimental results on the arXiv benchma
rk continue to demonstrate its superior performance.
****
Probabilistic Generating Circuits
Honghua Zhang, Brendan Juba, Guy Van Den Broeck
Generating functions, which are widely used in combinatorics and probability the
ory, encode function values into the coefficients of a polynomial. In this paper
, we explore their use as a tractable probabilistic model, and propose probabili
stic generating circuits (PGCs) for their efficient representation. PGCs are str
ictly more expressive efficient than many existing tractable probabilistic model
s, including determinantal point processes (DPPs), probabilistic circuits (PCs)
such as sum-product networks, and tractable graphical models. We contend that PG
Cs are not just a theoretical framework that unifies vastly different existing m
odels, but also show great potential in modeling realistic data. We exhibit a si
mple class of PGCs that are not trivially subsumed by simple combinations of PCs

and DPPs, and obtain competitive performance on a suite of density estimation benchmarks. We also highlight PGCs' connection to the theory of strongly Rayleigh distributions.
****
PAPRIKA: Private Online False Discovery Rate Control
Wanrong Zhang, Gautam Kamath, Rachel Cummings
In hypothesis testing, a \emph{false discovery} occurs when a hypothesis is incorrectly rejected due to noise in the sample. When adaptively testing multiple hypotheses, the probability of a false discovery increases as more tests are performed. Thus the problem of \emph{False Discovery Rate (FDR) control} is to find a procedure for testing multiple hypotheses that accounts for this effect in determining the set of hypotheses to reject. The goal is to minimize the number (or fraction) of false discoveries, while maintaining a high true positive rate (i.e., correct discoveries). In this work, we study False Discovery Rate (FDR) control in multiple hypothesis testing under the constraint of differential privacy for the sample. Unlike previous work in this direction, we focus on the \emph{online setting}, meaning that a decision about each hypothesis must be made immediately after the test is performed, rather than waiting for the output of all tests as in the offline setting. We provide new private algorithms based on state-of-the-art results in non-private online FDR control. Our algorithms have strong provable guarantees for privacy and statistical performance as measured by FDR and power. We also provide experimental results to demonstrate the efficacy of our algorithms in a variety of data environments.
****
Learning from Noisy Labels with No Change to the Training Process
Mingyuan Zhang, Jane Lee, Shivani Agarwal
There has been much interest in recent years in developing learning algorithms that can learn accurate classifiers from data with noisy labels. A widely-studied noise model is that of \emph{class-conditional noise} (CCN), wherein a label y is flipped to a label $\tilde{y}$ with some associated noise probability that depends on both y and $\tilde{y}$. In the multiclass setting, all previously proposed algorithms under the CCN model involve changing the training process, by introducing a 'noise-correction' to the surrogate loss to be minimized over the noisy training examples. In this paper, we show that this is really unnecessary: one can simply perform class probability estimation (CPE) on the noisy examples, e.g. using a standard (multiclass) logistic regression algorithm, and then apply noise-correction only in the final prediction step. This means that the training algorithm itself does not need any change, and one can simply use standard off-the-shelf implementations with no modification to the code for training. Our approach can handle general multiclass loss matrices, including the usual 0-1 loss but also other losses such as those used for ordinal regression problems. We also provide a quantitative regret transfer bound, which bounds the target regret on the true distribution in terms of the CPE regret on the noisy distribution; in doing so, we extend the notion of strong properness introduced for binary losses by Agarwal (2014) to the multiclass case. Our bound suggests that the sample complexity of learning under CCN increases as the noise matrix approaches singularity. We also provide fixes and potential improvements for noise estimation methods that involve computing anchor points. Our experiments confirm our theoretical findings.
****
Progressive-Scale Boundary Blackbox Attack via Projective Gradient Estimation
Jiawei Zhang, Linyi Li, Huichen Li, Xiaolu Zhang, Shuang Yang, Bo Li
Boundary based blackbox attack has been recognized as practical and effective, given that an attacker only needs to access the final model prediction. However, the query efficiency of it is in general high especially for high dimensional image data. In this paper, we show that such efficiency highly depends on the scale at which the attack is applied, and attacking at the optimal scale significantly improves the efficiency. In particular, we propose a theoretical framework to analyze and show three key characteristics to improve the query efficiency. We prove that there exists an optimal scale for projective gradient estimation. Our

framework also explains the satisfactory performance achieved by existing boundary black-box attacks. Based on our theoretical framework, we propose Progressive-Scale enabled projective Boundary Attack (PSBA) to improve the query efficiency via progressive scaling techniques. In particular, we employ Progressive-GAN to optimize the scale of projections, which we call PSBA-PGAN. We evaluate our approach on both spatial and frequency scales. Extensive experiments on MNIST, CIFAR-10, CelebA, and ImageNet against different models including a real-world face recognition API show that PSBA-PGAN significantly outperforms existing baseline attacks in terms of query efficiency and attack success rate. We also observe relatively stable optimal scales for different models and datasets. The code is publicly available at https://github.com/AI-secure/PSBA.
****
FOP: Factorizing Optimal Joint Policy of Maximum-Entropy Multi-Agent Reinforcement Learning
Tianhao Zhang, Yueheng Li, Chen Wang, Guangming Xie, Zongqing Lu
Value decomposition recently injects vigorous vitality into multi-agent actor-critic methods. However, existing decomposed actor-critic methods cannot guarantee the convergence of global optimum. In this paper, we present a novel multi-agent actor-critic method, FOP, which can factorize the optimal joint policy induced by maximum-entropy multi-agent reinforcement learning (MARL) into individual policies. Theoretically, we prove that factorized individual policies of FOP converge to the global optimum. Empirically, in the well-known matrix game and differential game, we verify that FOP can converge to the global optimum for both discrete and continuous action spaces. We also evaluate FOP on a set of StarCraft II micromanagement tasks, and demonstrate that FOP substantially outperforms state-of-the-art decomposed value-based and actor-critic methods.
****
Learning Noise Transition Matrix from Only Noisy Labels via Total Variation Regularization
Yivan Zhang, Gang Niu, Masashi Sugiyama
Many weakly supervised classification methods employ a noise transition matrix to capture the class-conditional label corruption. To estimate the transition matrix from noisy data, existing methods often need to estimate the noisy class-posterior, which could be unreliable due to the overconfidence of neural networks. In this work, we propose a theoretically grounded method that can estimate the noise transition matrix and learn a classifier simultaneously, without relying on the error-prone noisy class-posterior estimation. Concretely, inspired by the characteristics of the stochastic label corruption process, we propose total variation regularization, which encourages the predicted probabilities to be more distinguishable from each other. Under mild assumptions, the proposed method yields a consistent estimator of the transition matrix. We show the effectiveness of the proposed method through experiments on benchmark and real-world datasets.
****
Quantile Bandits for Best Arms Identification
Mengyan Zhang, Cheng Soon Ong
We consider a variant of the best arm identification task in stochastic multi-armed bandits. Motivated by risk-averse decision-making problems, our goal is to identify a set of m arms with the highest τ-quantile values within a fixed budget. We prove asymmetric two-sided concentration inequalities for order statistics and quantiles of random variables that have non-decreasing hazard rate, which may be of independent interest. With these inequalities, we analyse a quantile version of Successive Accepts and Rejects (Q-SAR). We derive an upper bound for the probability of arm misidentification, the first justification of a quantile based algorithm for fixed budget multiple best arms identification. We show illustrative experiments for best arm identification.
****
Towards Better Robust Generalization with Shift Consistency Regularization
Shufei Zhang, Zhuang Qian, Kaizhu Huang, Qiufeng Wang, Rui Zhang, Xinping Yi
While adversarial training becomes one of the most promising defending approaches against adversarial attacks for deep neural networks, the conventional wisdom

through robust optimization may usually not guarantee good generalization for ro bustness. Concerning with robust generalization over unseen adversarial data, th is paper investigates adversarial training from a novel perspective of shift con sistency in latent space. We argue that the poor robust generalization of advers arial training is owing to the significantly dispersed latent representations ge nerated by training and test adversarial data, as the adversarial perturbations push the latent features of natural examples in the same class towards diverse d irections. This is underpinned by the theoretical analysis of the robust general ization gap, which is upper-bounded by the standard one over the natural data an d a term of feature inconsistent shift caused by adversarial perturbation {–} a measure of latent dispersion. Towards better robust generalization, we propose a new regularization method {–} shift consistency regularization (SCR) {–} to ste er the same-class latent features of both natural and adversarial data into a co mmon direction during adversarial training. The effectiveness of SCR in adversar ial training is evaluated through extensive experiments over different datasets, such as CIFAR-10, CIFAR-100, and SVHN, against several competitive methods.
****
On-Policy Deep Reinforcement Learning for the Average-Reward Criterion
Yiming Zhang, Keith W Ross
We develop theory and algorithms for average-reward on-policy Reinforcement Lear ning (RL). We first consider bounding the difference of the long-term average re ward for two policies. We show that previous work based on the discounted return (Schulman et al. 2015, Achiam et al. 2017) results in a non-meaningful lower bo und in the average reward setting. By addressing the average-reward criterion di rectly, we then derive a novel bound which depends on the average divergence bet ween the policies and on Kemeny’s constant. Based on this bound, we develop an i terative procedure which produces a sequence of monotonically improved policies for the average reward criterion. This iterative procedure can then be combined with classic Deep Reinforcement Learning (DRL) methods, resulting in practical D RL algorithms that target the long-run average reward criterion. In particular, we demonstrate that Average-Reward TRPO (ATRPO), which adapts the on-policy TRPO algorithm to the average-reward criterion, significantly outperforms TRPO in th e most challenging MuJuCo environments.
****
Differentiable Dynamic Quantization with Mixed Precision and Adaptive Resolution
Zhaoyang Zhang, Wenqi Shao, Jinwei Gu, Xiaogang Wang, Ping Luo
Model quantization is challenging due to many tedious hyper-parameters such as p recision (bitwidth), dynamic range (minimum and maximum discrete values) and ste psize (interval between discrete values). Unlike prior arts that carefully tune these values, we present a fully differentiable approach to learn all of them, n amed Differentiable Dynamic Quantization (DDQ), which has several benefits. (1) DDQ is able to quantize challenging lightweight architectures like MobileNets, w here different layers prefer different quantization parameters. (2) DDQ is hardw are-friendly and can be easily implemented using low-precision matrix-vector mul tiplication, making it capable in many hardware such as ARM. (3) Extensive exper iments show that DDQ outperforms prior arts on many networks and benchmarks, esp ecially when models are already efficient and compact. e.g., DDQ is the first ap proach that achieves lossless 4-bit quantization for MobileNetV2 on ImageNet.
****
iDARTS: Differentiable Architecture Search with Stochastic Implicit Gradients
Miao Zhang, Steven W. Su, Shirui Pan, Xiaojun Chang, Ehsan M Abbasnejad, Reza Ha ffari
Differentiable ARchiTecture Search(DARTS) has recently become the mainstream in the neural architecture search (NAS) due to its efficiency and simplicity. With a gradient-based bi-level optimization, DARTS alternately optimizes the inner mo del weights and the outer architecture parameter in a weight-sharing supernet. A key challenge to the scalability and quality of the learned architectures is th e need for differentiating through the inner-loop optimisation. While much has b een discussed about several potentially fatal factors in DARTS, the architecture gradient, a.k.a. hypergradient, has received less attention. In this paper, we

tackle the hypergradient computation in DARTS based on the implicit function theorem, making it only depends on the obtained solution to the inner-loop optimization and agnostic to the optimization path. To further reduce the computational requirements, we formulate a stochastic hypergradient approximation for differentiable NAS, and theoretically show that the architecture optimization with the proposed method is expected to converge to a stationary point. Comprehensive experiments on two NAS benchmark search spaces and the common NAS search space verify the effectiveness of our proposed method. It leads to architectures outperforming, with large margins, those learned by the baseline methods.
****

Deep Coherent Exploration for Continuous Control
Yijie Zhang, Herke Van Hoof
In policy search methods for reinforcement learning (RL), exploration is often performed by injecting noise either in action space at each step independently or in parameter space over each full trajectory. In prior work, it has been shown that with linear policies, a more balanced trade-off between these two exploration strategies is beneficial. However, that method did not scale to policies using deep neural networks. In this paper, we introduce deep coherent exploration, a general and scalable exploration framework for deep RL algorithms for continuous control, that generalizes step-based and trajectory-based exploration. This framework models the last layer parameters of the policy network as latent variables and uses a recursive inference step within the policy update to handle these latent variables in a scalable manner. We find that deep coherent exploration improves the speed and stability of learning of A2C, PPO, and SAC on several continuous control tasks.
****

Average-Reward Off-Policy Policy Evaluation with Function Approximation
Shangtong Zhang, Yi Wan, Richard S Sutton, Shimon Whiteson
We consider off-policy policy evaluation with function approximation (FA) in average-reward MDPs, where the goal is to estimate both the reward rate and the differential value function. For this problem, bootstrapping is necessary and, along with off-policy learning and FA, results in the deadly triad (Sutton & Barto, 2018). To address the deadly triad, we propose two novel algorithms, reproducing the celebrated success of Gradient TD algorithms in the average-reward setting. In terms of estimating the differential value function, the algorithms are the first convergent off-policy linear function approximation algorithms. In terms of estimating the reward rate, the algorithms are the first convergent off-policy linear function approximation algorithms that do not require estimating the density ratio. We demonstrate empirically the advantage of the proposed algorithms, as well as their nonlinear variants, over a competitive density-ratio-based approach, in a simple domain as well as challenging robot simulation tasks.
****

Matrix Sketching for Secure Collaborative Machine Learning
Mengjiao Zhang, Shusen Wang
Collaborative learning allows participants to jointly train a model without data sharing. To update the model parameters, the central server broadcasts model parameters to the clients, and the clients send updating directions such as gradients to the server. While data do not leave a client device, the communicated gradients and parameters will leak a client’s privacy. Attacks that infer clients’ privacy from gradients and parameters have been developed by prior work. Simple defenses such as dropout and differential privacy either fail to defend the attacks or seriously hurt test accuracy. We propose a practical defense which we call Double-Blind Collaborative Learning (DBCL). The high-level idea is to apply random matrix sketching to the parameters (aka weights) and re-generate random sketching after each iteration. DBCL prevents clients from conducting gradient-based privacy inferences which are the most effective attacks. DBCL works because from the attacker’s perspective, sketching is effectively random noise that outweighs the signal. Notably, DBCL does not much increase computation and communication costs and does not hurt test accuracy at all.
****

MetaCURE: Meta Reinforcement Learning with Empowerment-Driven Exploration
Jin Zhang, Jianhao Wang, Hao Hu, Tong Chen, Yingfeng Chen, Changjie Fan, Chongji e Zhang
Meta reinforcement learning (meta-RL) extracts knowledge from previous tasks and achieves fast adaptation to new tasks. Despite recent progress, efficient exploration in meta-RL remains a key challenge in sparse-reward tasks, as it requires quickly finding informative task-relevant experiences in both meta-training and adaptation. To address this challenge, we explicitly model an exploration policy learning problem for meta-RL, which is separated from exploitation policy learning, and introduce a novel empowerment-driven exploration objective, which aims to maximize information gain for task identification. We derive a corresponding intrinsic reward and develop a new off-policy meta-RL framework, which efficiently learns separate context-aware exploration and exploitation policies by sharing the knowledge of task inference. Experimental evaluation shows that our meta-RL method significantly outperforms state-of-the-art baselines on various sparse-reward MuJoCo locomotion tasks and more complex sparse-reward Meta-World tasks.
****
World Model as a Graph: Learning Latent Landmarks for Planning
Lunjun Zhang, Ge Yang, Bradly C Stadie
Planning, the ability to analyze the structure of a problem in the large and decompose it into interrelated subproblems, is a hallmark of human intelligence. While deep reinforcement learning (RL) has shown great promise for solving relatively straightforward control tasks, it remains an open problem how to best incorporate planning into existing deep RL paradigms to handle increasingly complex environments. One prominent framework, Model-Based RL, learns a world model and plans using step-by-step virtual rollouts. This type of world model quickly diverges from reality when the planning horizon increases, thus struggling at long-horizon planning. How can we learn world models that endow agents with the ability to do temporally extended reasoning? In this work, we propose to learn graph-structured world models composed of sparse, multi-step transitions. We devise a novel algorithm to learn latent landmarks that are scattered (in terms of reachability) across the goal space as the nodes on the graph. In this same graph, the edges are the reachability estimates distilled from Q-functions. On a variety of high-dimensional continuous control tasks ranging from robotic manipulation to navigation, we demonstrate that our method, named L3P, significantly outperforms prior work, and is oftentimes the only method capable of leveraging both the robustness of model-free RL and generalization of graph-search algorithms. We believe our work is an important step towards scalable planning in reinforcement learning.
****
Breaking the Deadly Triad with a Target Network
Shangtong Zhang, Hengshuai Yao, Shimon Whiteson
The deadly triad refers to the instability of a reinforcement learning algorithm when it employs off-policy learning, function approximation, and bootstrapping simultaneously. In this paper, we investigate the target network as a tool for breaking the deadly triad, providing theoretical support for the conventional wisdom that a target network stabilizes training. We first propose and analyze a novel target network update rule which augments the commonly used Polyak-averaging style update with two projections. We then apply the target network and ridge regularization in several divergent algorithms and show their convergence to regularized TD fixed points. Those algorithms are off-policy with linear function approximation and bootstrapping, spanning both policy evaluation and control, as well as both discounted and average-reward settings. In particular, we provide the first convergent linear Q-learning algorithms under nonrestrictive and changing behavior policies without bi-level optimization.
****
Multiscale Invertible Generative Networks for High-Dimensional Bayesian Inference
Shumao Zhang, Pengchuan Zhang, Thomas Y Hou
We propose a Multiscale Invertible Generative Network (MsIGN) and associated tra

ining algorithm that leverages multiscale structure to solve high-dimensional Ba yesian inference. To address the curse of dimensionality, MsIGN exploits the low -dimensional nature of the posterior, and generates samples from coarse to fine scale (low to high dimension) by iteratively upsampling and refining samples. Ms IGN is trained in a multi-stage manner to minimize the Jeffreys divergence, whic h avoids mode dropping in high-dimensional cases. On two high-dimensional Bayesi an inverse problems, we show superior performance of MsIGN over previous approac hes in posterior approximation and multiple mode capture. On the natural image s ynthesis task, MsIGN achieves superior performance in bits-per-dimension over ba seline models and yields great interpret-ability of its neurons in intermediate layers.
****
Meta Learning for Support Recovery in High-dimensional Precision Matrix Estimati on
Qian Zhang, Yilin Zheng, Jean Honorio
In this paper, we study meta learning for support (i.e., the set of non-zero ent ries) recovery in high-dimensional precision matrix estimation where we reduce t he sufficient sample complexity in a novel task with the information learned fro m other auxiliary tasks. In our setup, each task has a different random true pre cision matrix, each with a possibly different support. We assume that the union of the supports of all the true precision matrices (i.e., the true support union) is small in size. We propose to pool all the samples from different tasks, and \emph{improperly} estimate a single precision matrix by minimizing the ℓ_1 -regularized log-determinant Bregman divergence. We show that with high probabil ity, the support of the \emph{improperly} estimated single precision matrix is e qual to the true support union, provided a sufficient number of samples per task $n \in O((\log N)/K)$, for N-dimensional vectors and K tasks. That is, one requires less samples per task when more tasks are available. We prove a matchin g information-theoretic lower bound for the necessary number of samples, which i s $n \in \Omega((\log N)/K)$, and thus, our algorithm is minimax optimal. Then f or the novel task, we prove that the minimization of the ℓ_1-regularized lo g-determinant Bregman divergence with the additional constraint that the support is a subset of the estimated support union could reduce the sufficient sample c omplexity of successful support recovery to $O(\log(|S_{\text{off}}|))$ where $|S_{\text{off}}|$ is the number of off-diagonal elements in the support union and is much less than N for sparse matrices. We also prove a matching information -theoretic lower bound of $\Omega(\log(|S_{\text{off}}|))$ for the necessary num ber of samples.
****
Model-Free Reinforcement Learning: from Clipped Pseudo-Regret to Sample Complexi ty
Zihan Zhang, Yuan Zhou, Xiangyang Ji
In this paper we consider the problem of learning an ϵ-optimal policy f or a discounted Markov Decision Process (MDP). Given an MDP with S states, A actions, the discount factor $\gamma \in (0,1)$, and an approximation threshold $\epsilon > 0$, we provide a model-free algorithm to learn an ϵ-optima l policy with sample complexity $\tilde{O}(\frac{SA\ln(1/p)}{\epsilon^2(1-\gamma)^{5.5}})$ \footnote{In this work, the notation $\tilde{O}(\cdot)$ hides poly-lo garithmic factors of $S,A,1/(1-\gamma)$, and $1/\epsilon$.} and success probabil ity $(1-p)$. For small enough ϵ, we show an improved algorithm with sam ple complexity $\tilde{O}(\frac{SA\ln(1/p)}{\epsilon^2(1-\gamma)^{3}})$. While t he first bound improves upon all known model-free algorithms and model-based one s with tight dependence on S, our second algorithm beats all known sample comp lexity bounds and matches the information theoretic lower bound up to logarithmi c factors.
****
Learning to Rehearse in Long Sequence Memorization
Zhu Zhang, Chang Zhou, Jianxin Ma, Zhijie Lin, Jingren Zhou, Hongxia Yang, Zhou Zhao
Existing reasoning tasks often have an important assumption that the input conte

nts can be always accessed while reasoning, requiring unlimited storage resource s and suffering from severe time delay on long sequences. To achieve efficient r easoning on long sequences with limited storage resources, memory augmented neur al networks introduce a human-like write-read memory to compress and memorize th e long input sequence in one pass, trying to answer subsequent queries only base d on the memory. But they have two serious drawbacks: 1) they continually update the memory from current information and inevitably forget the early contents; 2) they do not distinguish what information is important and treat all contents e qually. In this paper, we propose the Rehearsal Memory (RM) to enhance long-sequ ence memorization by self-supervised rehearsal with a history sampler. To allevi ate the gradual forgetting of early information, we design self-supervised rehea rsal training with recollection and familiarity tasks. Further, we design a hist ory sampler to select informative fragments for rehearsal training, making the m emory focus on the crucial information. We evaluate the performance of our rehea rsal memory by the synthetic bAbI task and several downstream tasks, including t ext/video question answering and recommendation on long sequences.
****
Dataset Condensation with Differentiable Siamese Augmentation
Bo Zhao, Hakan Bilen
In many machine learning problems, large-scale datasets have become the de-facto standard to train state-of-the-art deep networks at the price of heavy computat ion load. In this paper, we focus on condensing large training sets into signifi cantly smaller synthetic sets which can be used to train deep neural networks fr om scratch with minimum drop in performance. Inspired from the recent training s et synthesis methods, we propose Differentiable Siamese Augmentation that enable s effective use of data augmentation to synthesize more informative synthetic im ages and thus achieves better performance when training networks with augmentati ons. Experiments on multiple image classification benchmarks demonstrate that th e proposed method obtains substantial gains over the state-of-the-art, 7% improv ements on CIFAR10 and CIFAR100 datasets. We show with only less than 1% data tha t our method achieves 99.6%, 94.9%, 88.5%, 71.5% relative performance on MNIST, FashionMNIST, SVHN, CIFAR10 respectively. We also explore the use of our method in continual learning and neural architecture search, and show promising results .
****
Joining datasets via data augmentation in the label space for neural networks
Junbo Zhao, Mingfeng Ou, Linji Xue, Yunkai Cui, Sai Wu, Gang Chen
Most, if not all, modern deep learning systems restrict themselves to a single d ataset for neural network training and inference. In this article, we are intere sted in systematic ways to join datasets that are made of similar purposes. Unli ke previous published works that ubiquitously conduct the dataset joining in the uninterpretable latent vectorial space, the core to our method is an augmentati on procedure in the label space. The primary challenge to address the label spac e for dataset joining is the discrepancy between labels: non-overlapping label a nnotation sets, different labeling granularity or hierarchy and etc. Notably we propose a new technique leveraging artificially created knowledge graph, recurre nt neural networks and policy gradient that successfully achieve the dataset joi ning in the label space. Empirical results on both image and text classification justify the validity of our approach.
****
Calibrate Before Use: Improving Few-shot Performance of Language Models
Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, Sameer Singh
GPT-3 can perform numerous tasks when provided a natural language prompt that co ntains a few training examples. We show that this type of few-shot learning can be unstable: the choice of prompt format, training examples, and even the order of the examples can cause accuracy to vary from near chance to near state-of-the -art. We demonstrate that this instability arises from the bias of language mode ls towards predicting certain answers, e.g., those that are placed near the end of the prompt or are common in the pre-training data. To mitigate this, we first estimate the model’s bias towards each answer by asking for its prediction when

given a training prompt and a content-free test input such as "N/A". We then fit calibration parameters that cause the prediction for this input to be uniform across answers. On a diverse set of tasks, this contextual calibration procedure substantially improves GPT-3 and GPT-2's accuracy (up to 30.0% absolute) across different choices of the prompt, while also making learning considerably more stable.
****
Few-Shot Neural Architecture Search
Yiyang Zhao, Linnan Wang, Yuandong Tian, Rodrigo Fonseca, Tian Guo
Efficient evaluation of a network architecture drawn from a large search space remains a key challenge in Neural Architecture Search (NAS). Vanilla NAS evaluates each architecture by training from scratch, which gives the true performance but is extremely time-consuming. Recently, one-shot NAS substantially reduces the computation cost by training only one supernetwork, a.k.a. supernet, to approximate the performance of every architecture in the search space via weight-sharing. However, the performance estimation can be very inaccurate due to the co-adaption among operations. In this paper, we propose few-shot NAS that uses multiple supernetworks, called sub-supernet, each covering different regions of the search space to alleviate the undesired co-adaption. Compared to one-shot NAS, few-shot NAS improves the accuracy of architecture evaluation with a small increase of evaluation cost. With only up to 7 sub-supernets, few-shot NAS establishes new SoTAs: on ImageNet, it finds models that reach 80.5% top-1 accuracy at 600 MB FLOPS and 77.5% top-1 accuracy at 238 MFLOPS; on CIFAR10, it reaches 98.72% top-1 accuracy without using extra data or transfer learning. In Auto-GAN, few-shot NAS outperforms the previously published results by up to 20%. Extensive experiments show that few-shot NAS significantly improves various one-shot methods, including 4 gradient-based and 6 search-based methods on 3 different tasks in NasBench-201 and NasBench1-shot-1.
****
Expressive 1-Lipschitz Neural Networks for Robust Multiple Graph Learning against Adversarial Attacks
Xin Zhao, Zeru Zhang, Zijie Zhang, Lingfei Wu, Jiayin Jin, Yang Zhou, Ruoming Jin, Dejing Dou, Da Yan
Recent findings have shown multiple graph learning models, such as graph classification and graph matching, are highly vulnerable to adversarial attacks, i.e. small input perturbations in graph structures and node attributes can cause the model failures. Existing defense techniques often defend specific attacks on particular multiple graph learning tasks. This paper proposes an attack-agnostic graph-adaptive 1-Lipschitz neural network, ERNN, for improving the robustness of deep multiple graph learning while achieving remarkable expressive power. A K_l-Lipschitz Weibull activation function is designed to enforce the gradient norm as K_l at layer l. The nearest matrix orthogonalization and polar decomposition techniques are utilized to constraint the weight norm as 1/K_l and make the norm-constrained weight close to the original weight. The theoretical analysis is conducted to derive lower and upper bounds of feasible K_l under the 1-Lipschitz constraint. The combination of norm-constrained weight and activation function leads to the 1-Lipschitz neural network for expressive and robust multiple graph learning.
****
Fused Acoustic and Text Encoding for Multimodal Bilingual Pretraining and Speech Translation
Renjie Zheng, Junkun Chen, Mingbo Ma, Liang Huang
Recently, representation learning for text and speech has successfully improved many language related tasks. However, all existing methods suffer from two limitations: (a) they only learn from one input modality, while a unified representation for both speech and text is needed by tasks such as end-to-end speech translation, and as a result, (b) they can not exploit various large-scale text and speech data and their performance is limited by the scarcity of parallel speech translation data. To address these problems, we propose a Fused Acoustic and Text Masked Language Model (FAT-MLM) which jointly learns a unified representation fo

r both acoustic and text input from various types of corpora including parallel data for speech recognition and machine translation, and even pure speech and text data. Within this cross-modal representation learning framework, we further present an end-to-end model for Fused Acoustic and Text Speech Translation (FAT-ST). Experiments on three translation directions show that by fine-tuning from FAT-MLM, our proposed speech translation models substantially improve translation quality by up to +5.9 BLEU.
****
Two Heads are Better Than One: Hypergraph-Enhanced Graph Reasoning for Visual Event Ratiocination
Wenbo Zheng, Lan Yan, Chao Gou, Fei-Yue Wang
Even with a still image, humans can ratiocinate various visual cause-and-effect descriptions before, at present, and after, as well as beyond the given image. However, it is challenging for models to achieve such task–the visual event ratiocination, owing to the limitations of time and space. To this end, we propose a novel multi-modal model, Hypergraph-Enhanced Graph Reasoning. First it represents the contents from the same modality as a semantic graph and mines the intra-modality relationship, therefore breaking the limitations in the spatial domain. Then, we introduce the Graph Self-Attention Enhancement. On the one hand, this enables semantic graph representations from different modalities to enhance each other and captures the inter-modality relationship along the line. On the other hand, it utilizes our built multi-modal hypergraphs in different moments to boost individual semantic graph representations, and breaks the limitations in the temporal domain. Our method illustrates the case of "two heads are better than one" in the sense that semantic graph representations with the help of the proposed enhancement mechanism are more robust than those without. Finally, we re-project these representations and leverage their outcomes to generate textual cause-and-effect descriptions. Experimental results show that our model achieves significantly higher performance in comparison with other state-of-the-arts.
****
How Framelets Enhance Graph Neural Networks
Xuebin Zheng, Bingxin Zhou, Junbin Gao, Yuguang Wang, Pietro Lió, Ming Li, Guido Montufar
This paper presents a new approach for assembling graph neural networks based on framelet transforms. The latter provides a multi-scale representation for graph-structured data. We decompose an input graph into low-pass and high-pass frequencies coefficients for network training, which then defines a framelet-based graph convolution. The framelet decomposition naturally induces a graph pooling strategy by aggregating the graph feature into low-pass and high-pass spectra, which considers both the feature values and geometry of the graph data and conserves the total information. The graph neural networks with the proposed framelet convolution and pooling achieve state-of-the-art performance in many node and graph prediction tasks. Moreover, we propose shrinkage as a new activation for the framelet convolution, which thresholds high-frequency information at different scales. Compared to ReLU, shrinkage activation improves model performance on denoising and signal compression: noises in both node and structure can be significantly reduced by accurately cutting off the high-pass coefficients from framelet decomposition, and the signal can be compressed to less than half its original size with well-preserved prediction performance.
****
Probabilistic Sequential Shrinking: A Best Arm Identification Algorithm for Stochastic Bandits with Corruptions
Zixin Zhong, Wang Chi Cheung, Vincent Tan
We consider a best arm identification (BAI) problem for stochastic bandits with adversarial corruptions in the fixed-budget setting of T steps. We design a novel randomized algorithm, Probabilistic Sequential Shrinking(u) (PSS(u)), which is agnostic to the amount of corruptions. When the amount of corruptions per step (CPS) is below a threshold, PSS(u) identifies the best arm or item with probability tending to 1 as T{\rightarrow}∞. Otherwise, the optimality gap of the identified item degrades gracefully with the CPS.We argue that such a bifurcati

on is necessary. In PSS(u), the parameter u serves to balance between the optima
lity gap and success probability. The injection of randomization is shown to be
essential to mitigate the impact of corruptions. To demonstrate this, we design
two attack strategies that are applicable to any algorithm. We apply one of them
to a deterministic analogue of PSS(u) known as Successive Halving (SH) by Karni
n et al. (2013). The attack strategy results in a high failure probability for S
H, but PSS(u) remains robust. In the absence of corruptions, PSS(2)’s performanc
e guarantee matches SH’s. We show that when the CPS is sufficiently large, no al
gorithm can achieve a BAI probability tending to 1 as T{\rightarrow}∞. Nu
merical experiments corroborate our theoretical findings.
****
Towards Distraction-Robust Active Visual Tracking
Fangwei Zhong, Peng Sun, Wenhan Luo, Tingyun Yan, Yizhou Wang
In active visual tracking, it is notoriously difficult when distracting objects
appear, as distractors often mislead the tracker by occluding the target or brin
ging a confusing appearance. To address this issue, we propose a mixed cooperati
ve-competitive multi-agent game, where a target and multiple distractors form a
collaborative team to play against a tracker and make it fail to follow. Through
learning in our game, diverse distracting behaviors of the distractors naturall
y emerge, thereby exposing the tracker’s weakness, which helps enhance the distr
action-robustness of the tracker. For effective learning, we then present a bunc
h of practical methods, including a reward function for distractors, a cross-mod
al teacher-student learning strategy, and a recurrent attention mechanism for th
e tracker. The experimental results show that our tracker performs desired distr
action-robust active visual tracking and can be well generalized to unseen envir
onments. We also show that the multi-agent game can be used to adversarially tes
t the robustness of trackers.
****
Provably Efficient Reinforcement Learning for Discounted MDPs with Feature Mappi
ng
Dongruo Zhou, Jiafan He, Quanquan Gu
Modern tasks in reinforcement learning have large state and action spaces. To de
al with them efficiently, one often uses predefined feature mapping to represent
states and actions in a low dimensional space. In this paper, we study reinforc
ement learning for discounted Markov Decision Processes (MDPs), where the transi
tion kernel can be parameterized as a linear function of certain feature mapping
. We propose a novel algorithm which makes use of the feature mapping and obtain
s a $\tilde O(d\sqrt{T}/(1-\gamma)^2)$ regret, where d is the dimension of the
feature space, T is the time horizon and γ is the discount factor of t
he MDP. To the best of our knowledge, this is the first polynomial regret bound
without accessing a generative model or making strong assumptions such as ergodi
city of the MDP. By constructing a special class of MDPs, we also show that for
any algorithms, the regret is lower bounded by $\Omega(d\sqrt{T}/(1-\gamma)^{1.5
})$. Our upper and lower bound results together suggest that the proposed reinfo
rcement learning algorithm is near-optimal up to a $(1-\gamma)^{-0.5}$ factor.
****
Amortized Conditional Normalized Maximum Likelihood: Reliable Out of Distributio
n Uncertainty Estimation
Aurick Zhou, Sergey Levine
While deep neural networks provide good performance for a range of challenging t
asks, calibration and uncertainty estimation remain major challenges, especially
under distribution shift. In this paper, we propose the amortized conditional n
ormalized maximum likelihood (ACNML) method as a scalable general-purpose approa
ch for uncertainty estimation, calibration, and out-of-distribution robustness w
ith deep networks. Our algorithm builds on the conditional normalized maximum li
kelihood (CNML) coding scheme, which has minimax optimal properties according to
the minimum description length principle, but is computationally intractable to
evaluate exactly for all but the simplest of model classes. We propose to use a
pproximate Bayesian inference technqiues to produce a tractable approximation to
the CNML distribution. Our approach can be combined with any approximate infere

nce algorithm that provides tractable posterior densities over model parameters. We demonstrate that ACNML compares favorably to a number of prior techniques for uncertainty estimation in terms of calibration when faced with distribution shift.
****
Optimal Estimation of High Dimensional Smooth Additive Function Based on Noisy Observations
Fan Zhou, Ping Li
Given $\bx_j = \btheta + \bepsilon_j$, $j=1,...,n$ where $\btheta \in \RR^d$ is an unknown parameter and $\bepsilon_j$ are i.i.d. Gaussian noise vectors, we study the estimation of $f(\btheta)$ for a given smooth function $f:\RR^d \rightarrow \RR$ equipped with an additive structure. We inherit the idea from a recent work which introduced an effective bias reduction technique through iterative bootstrap and derive a bias-reducing estimator. By establishing its normal approximation results, we show that the proposed estimator can achieve asymptotic normality with a looser constraint on smoothness compared with general smooth function due to the additive structure. Such results further imply that the proposed estimator is asymptotically efficient. Both upper and lower bounds on mean squared error are proved which shows the proposed estimator is minimax optimal for the smooth class considered. Numerical simulation results are presented to validate our analysis and show its superior performance of the proposed estimator over the plug-in approach in terms of bias reduction and building confidence intervals.
****
Incentivized Bandit Learning with Self-Reinforcing User Preferences
Tianchen Zhou, Jia Liu, Chaosheng Dong, Jingyuan Deng
In this paper, we investigate a new multi-armed bandit (MAB) online learning model that considers real-world phenomena in many recommender systems: (i) the learning agent cannot pull the arms by itself and thus has to offer rewards to users to incentivize arm-pulling indirectly; and (ii) if users with specific arm preferences are well rewarded, they induce a "self-reinforcing" effect in the sense that they will attract more users of similar arm preferences. Besides addressing the tradeoff of exploration and exploitation, another key feature of this new MAB model is to balance reward and incentivizing payment. The goal of the agent is to maximize the total reward over a fixed time horizon T with a low total payment. Our contributions in this paper are two-fold: (i) We propose a new MAB model with random arm selection that considers the relationship of users' self-reinforcing preferences and incentives; and (ii) We leverage the properties of a multi-color Polya urn with nonlinear feedback model to propose two MAB policies termed "At-Least-n Explore-Then-Commit" and "UCB-List". We prove that both policies achieve $O(log T)$ expected regret with $O(log T)$ expected payment over a time horizon T. We conduct numerical simulations to demonstrate and verify the performances of these two policies and study their robustness under various settings.
****
Towards Defending against Adversarial Examples via Attack-Invariant Features
Dawei Zhou, Tongliang Liu, Bo Han, Nannan Wang, Chunlei Peng, Xinbo Gao
Deep neural networks (DNNs) are vulnerable to adversarial noise. Their adversarial robustness can be improved by exploiting adversarial examples. However, given the continuously evolving attacks, models trained on seen types of adversarial examples generally cannot generalize well to unseen types of adversarial examples. To solve this problem, in this paper, we propose to remove adversarial noise by learning generalizable invariant features across attacks which maintain semantic classification information. Specifically, we introduce an adversarial feature learning mechanism to disentangle invariant features from adversarial noise. A normalization term has been proposed in the encoded space of the attack-invariant features to address the bias issue between the seen and unseen types of attacks. Empirical evaluations demonstrate that our method could provide better protection in comparison to previous state-of-the-art approaches, especially against unseen types of attacks and adaptive attacks.
****

Asymmetric Loss Functions for Learning with Noisy Labels
Xiong Zhou, Xianming Liu, Junjun Jiang, Xin Gao, Xiangyang Ji
Robust loss functions are essential for training deep neural networks with better generalization power in the presence of noisy labels. Symmetric loss functions are confirmed to be robust to label noise. However, the symmetric condition is overly restrictive. In this work, we propose a new class of loss functions, namely asymmetric loss functions, which are robust to learning from noisy labels for arbitrary noise type. Subsequently, we investigate general theoretical properties of asymmetric loss functions, including classification-calibration, excess risk bound, and noise-tolerance. Meanwhile, we introduce the asymmetry ratio to measure the asymmetry of a loss function, and the empirical results show that a higher ratio will provide better robustness. Moreover, we modify several common loss functions, and establish the necessary and sufficient conditions for them to be asymmetric. Experiments on benchmark datasets demonstrate that asymmetric loss functions can outperform state-of-the-art methods.
****
Examining and Combating Spurious Features under Distribution Shift
Chunting Zhou, Xuezhe Ma, Paul Michel, Graham Neubig
A central goal of machine learning is to learn robust representations that capture the fundamental relationship between inputs and output labels. However, minimizing training errors over finite or biased datasets results in models latching on to spurious correlations between the training input/output pairs that are not fundamental to the problem at hand. In this paper, we define and analyze robust and spurious representations using the information-theoretic concept of minimal sufficient statistics. We prove that even when there is only bias of the input distribution (i.e. covariate shift), models can still pick up spurious features from their training data. Group distributionally robust optimization (DRO) provides an effective tool to alleviate covariate shift by minimizing the worst-case training losses over a set of pre-defined groups. Inspired by our analysis, we demonstrate that group DRO can fail when groups do not directly account for various spurious correlations that occur in the data. To address this, we further propose to minimize the worst-case losses over a more flexible set of distributions that are defined on the joint distribution of groups and instances, instead of treating each group as a whole at optimization time. Through extensive experiments on one image and two language tasks, we show that our model is significantly more robust than comparable baselines under various partitions.
****
Sparse and Imperceptible Adversarial Attack via a Homotopy Algorithm
Mingkang Zhu, Tianlong Chen, Zhangyang Wang
Sparse adversarial attacks can fool deep neural networks (DNNs) by only perturbing a few pixels (regularized by ℓ_0 norm). Recent efforts combine it with another ℓ_∞ imperceptible on the perturbation magnitudes. The resultant sparse and imperceptible attacks are practically relevant, and indicate an even higher vulnerability of DNNs that we usually imagined. However, such attacks are more challenging to generate due to the optimization difficulty by coupling the ℓ_0 regularizer and box constraints with a non-convex objective. In this paper, we address this challenge by proposing a homotopy algorithm, to jointly tackle the sparsity and the perturbation bound in one unified framework. Each iteration, the main step of our algorithm is to optimize an ℓ_0-regularized adversarial loss, by leveraging the nonmonotone Accelerated Proximal Gradient Method (nmAPG) for nonconvex programming; it is followed by an ℓ_0 change control step, and an optional post-attack step designed to escape bad local minima. We also extend the algorithm to handling the structural sparsity regularizer. We extensively examine the effectiveness of our proposed \textbf{homotopy attack} for both targeted and non-targeted attack scenarios, on CIFAR-10 and ImageNet datasets. Compared to state-of-the-art methods, our homotopy attack leads to significantly fewer perturbations, e.g., reducing 42.91% on CIFAR-10 and 75.03% on ImageNet (average case, targeted attack), at similar maximal perturbation magnitudes, when still achieving 100% attack success rates. Our codes are available at: {\small\url{https://github.com/VITA-Group/SparseADV_Homotopy}}.

****
Data-Free Knowledge Distillation for Heterogeneous Federated Learning
Zhuangdi Zhu, Junyuan Hong, Jiayu Zhou
Federated Learning (FL) is a decentralized machine-learning paradigm, in which a global server iteratively averages the model parameters of local users without accessing their data. User heterogeneity has imposed significant challenges to FL, which can incur drifted global models that are slow to converge. Knowledge Distillation has recently emerged to tackle this issue, by refining the server model using aggregated knowledge from heterogeneous users, other than directly averaging their model parameters. This approach, however, depends on a proxy dataset, making it impractical unless such a prerequisite is satisfied. Moreover, the ensemble knowledge is not fully utilized to guide local model learning, which may in turn affect the quality of the aggregated model. Inspired by the prior art, we propose a data-free knowledge distillation approach to address heterogeneous FL, where the server learns a lightweight generator to ensemble user information in a data-free manner, which is then broadcasted to users, regulating local training using the learned knowledge as an inductive bias. Empirical studies powered by theoretical implications show that our approach facilitates FL with better generalization performance using fewer communication rounds, compared with the state-of-the-art.
****
Spectral vertex sparsifiers and pair-wise spanners over distributed graphs
Chunjiang Zhu, Qinqing Liu, Jinbo Bi
Graph sparsification is a powerful tool to approximate an arbitrary graph and has been used in machine learning over graphs. As real-world networks are becoming very large and naturally distributed, distributed graph sparsification has drawn considerable attention. In this work, we design communication-efficient distributed algorithms for constructing spectral vertex sparsifiers, which closely preserve effective resistance distances on a subset of vertices of interest in the original graphs, under the well-established message passing communication model. We prove that the communication cost approximates the lower bound with only a small gap. We further provide algorithms for constructing pair-wise spanners which approximate the shortest distances between each pair of vertices in a target set, instead of all pairs, and incur communication costs that are much smaller than those of existing algorithms in the message passing model. Experiments are performed to validate the communication efficiency of the proposed algorithms under the guarantee that the constructed sparsifiers have a good approximation quality.
****
Few-shot Language Coordination by Modeling Theory of Mind
Hao Zhu, Graham Neubig, Yonatan Bisk
No man is an island. Humans develop the ability to communicate with a large community by coordinating with different interlocutors within short conversations. This ability is largely understudied by the research on building neural language communicative agents. We study the task of few-shot language coordination: agents quickly adapting to their conversational partners’ language abilities. Different from current communicative agents trained with self-play, we in- investigate this more general paradigm by requiring the lead agent to coordinate with a population of agents each of whom has different linguistic abilities. This leads to a general agent able to quickly adapt to communicating with unseen agents in the population. Unlike prior work, success here requires the ability to model the partner’s beliefs, a vital component of human communication. Drawing inspiration from the study of theory-of-mind (ToM; Premack & Woodruff (1978)), we study the effect of the speaker explicitly modeling the listener’s mental state. Learning by communicating with a population, the speakers, as shown in our experiments, acquire the ability to learn to predict the reactions of their partner upon various messages on-the-fly. The speaker’s predictions for the future actions help it generate the best instructions in order to maximize communicative goal with message costs. To examine our hypothesis that the instructions generated with ToM modeling yield better communication per- performance, we employ our agents in bot

h a referential game and a language navigation task. Positive results from our e
xperiments also hint at the importance of explicitly modeling language acquisiti
on as a socio-pragmatic progress.
****
Clusterability as an Alternative to Anchor Points When Learning with Noisy Label
s
Zhaowei Zhu, Yiwen Song, Yang Liu
The label noise transition matrix, characterizing the probabilities of a trainin
g instance being wrongly annotated, is crucial to designing popular solutions to
 learning with noisy labels. Existing works heavily rely on finding “anchor poin
ts” or their approximates, defined as instances belonging to a particular class
almost surely. Nonetheless, finding anchor points remains a non-trivial task, an
d the estimation accuracy is also often throttled by the number of available anc
hor points. In this paper, we propose an alternative option to the above task. O
ur main contribution is the discovery of an efficient estimation procedure based
 on a clusterability condition. We prove that with clusterable representations o
f features, using up to third-order consensuses of noisy labels among neighbor r
epresentations is sufficient to estimate a unique transition matrix. Compared wi
th methods using anchor points, our approach uses substantially more instances a
nd benefits from a much better sample complexity. We demonstrate the estimation
accuracy and advantages of our estimates using both synthetic noisy labels (on C
IFAR-10/100) and real human-level noisy labels (on Clothing1M and our self-colle
cted human-annotated CIFAR-10). Our code and human-level noisy CIFAR-10 labels a
re available at https://github.com/UCSC-REAL/HOC.
****
Commutative Lie Group VAE for Disentanglement Learning
Xinqi Zhu, Chang Xu, Dacheng Tao
We view disentanglement learning as discovering an underlying structure that equ
ivariantly reflects the factorized variations shown in data. Traditionally, such
 a structure is fixed to be a vector space with data variations represented by t
ranslations along individual latent dimensions. We argue this simple structure i
s suboptimal since it requires the model to learn to discard the properties (e.g
. different scales of changes, different levels of abstractness) of data variati
ons, which is an extra work than equivariance learning. Instead, we propose to e
ncode the data variations with groups, a structure not only can equivariantly re
present variations, but can also be adaptively optimized to preserve the propert
ies of data variations. Considering it is hard to conduct training on group stru
ctures, we focus on Lie groups and adopt a parameterization using Lie algebra. B
ased on the parameterization, some disentanglement learning constraints are natu
rally derived. A simple model named Commutative Lie Group VAE is introduced to r
ealize the group-based disentanglement learning. Experiments show that our model
 can effectively learn disentangled representations without supervision, and can
 achieve state-of-the-art performance without extra constraints.
****
Accumulated Decoupled Learning with Gradient Staleness Mitigation for Convolutio
nal Neural Networks
Huiping Zhuang, Zhenyu Weng, Fulin Luo, Toh Kar-Ann, Haizhou Li, Zhiping Lin
Gradient staleness is a major side effect in decoupled learning when training co
nvolutional neural networks asynchronously. Existing methods that ignore this ef
fect might result in reduced generalization and even divergence. In this paper,
we propose an accumulated decoupled learning (ADL), which includes a module-wise
 gradient accumulation in order to mitigate the gradient staleness. Unlike prior
 arts ignoring the gradient staleness, we quantify the staleness in such a way t
hat its mitigation can be quantitatively visualized. As a new learning scheme, t
he proposed ADL is theoretically shown to converge to critical points in spite o
f its asynchronism. Extensive experiments on CIFAR-10 and ImageNet datasets are
conducted, demonstrating that ADL gives promising generalization results while t
he state-of-the-art methods experience reduced generalization and divergence. In
 addition, our ADL is shown to have the fastest training speed among the compare
d methods.

****
Demystifying Inductive Biases for (Beta-)VAE Based Architectures
Dominik Zietlow, Michal Rolinek, Georg Martius
The performance of Beta-Variational-Autoencoders and their variants on learning semantically meaningful, disentangled representations is unparalleled. On the other hand, there are theoretical arguments suggesting the impossibility of unsupervised disentanglement. In this work, we shed light on the inductive bias responsible for the success of VAE-based architectures. We show that in classical datasets the structure of variance, induced by the generating factors, is conveniently aligned with the latent directions fostered by the VAE objective. This builds the pivotal bias on which the disentangling abilities of VAEs rely. By small, elaborate perturbations of existing datasets, we hide the convenient correlation structure that is easily exploited by a variety of architectures. To demonstrate this, we construct modified versions of standard datasets in which (i) the generative factors are perfectly preserved; (ii) each image undergoes a mild transformation causing a small change of variance; (iii) the leading VAE-based disentanglement architectures fail to produce disentangled representations whilst the performance of a non-variational method remains unchanged.
****
Recovering AES Keys with a Deep Cold Boot Attack
Itamar Zimerman, Eliya Nachmani, Lior Wolf
Cold boot attacks inspect the corrupted random access memory soon after the power has been shut down. While most of the bits have been corrupted, many bits, at random locations, have not. Since the keys in many encryption schemes are being expanded in memory into longer keys with fixed redundancies, the keys can often be restored. In this work we combine a deep error correcting code technique together with a modified SAT solver scheme in order to apply the attack to AES keys. Even though AES consists Rijndael SBOX elements, that are specifically designed to be resistant to linear and differential cryptanalysis, our method provides a novel formalization of the AES key scheduling as a computational graph, which is implemented by neural message passing network. Our results show that our methods outperform the state of the art attack methods by a very large gap.
****
Learning Fair Policies in Decentralized Cooperative Multi-Agent Reinforcement Learning
Matthieu Zimmer, Claire Glanois, Umer Siddique, Paul Weng
We consider the problem of learning fair policies in (deep) cooperative multi-agent reinforcement learning (MARL). We formalize it in a principled way as the problem of optimizing a welfare function that explicitly encodes two important aspects of fairness: efficiency and equity. We provide a theoretical analysis of the convergence of policy gradient for this problem. As a solution method, we propose a novel neural network architecture, which is composed of two sub-networks specifically designed for taking into account these two aspects of fairness. In experiments, we demonstrate the importance of the two sub-networks for fair optimization. Our overall approach is general as it can accommodate any (sub)differentiable welfare function. Therefore, it is compatible with various notions of fairness that have been proposed in the literature (e.g., lexicographic maximin, generalized Gini social welfare function, proportional fairness). Our method is generic and can be implemented in various MARL settings: centralized training and decentralized execution, or fully decentralized. Finally, we experimentally validate our approach in various domains and show that it can perform much better than previous methods, both in terms of efficiency and equity.
****
Contrastive Learning Inverts the Data Generating Process
Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, Wieland Brendel
Contrastive learning has recently seen tremendous success in self-supervised learning. So far, however, it is largely unclear why the learned representations generalize so effectively to a large variety of downstream tasks. We here prove that feedforward models trained with objectives belonging to the commonly used Inf

oNCE family learn to implicitly invert the underlying generative model of the observed data. While the proofs make certain statistical assumptions about the generative model, we observe empirically that our findings hold even if these assumptions are severely violated. Our theory highlights a fundamental connection between contrastive learning, generative modeling, and nonlinear independent component analysis, thereby furthering our understanding of the learned representations as well as providing a theoretical foundation to derive more effective contrastive losses.
****
Exploration in Approximate Hyper-State Space for Meta Reinforcement Learning
Luisa M Zintgraf, Leo Feng, Cong Lu, Maximilian Igl, Kristian Hartikainen, Katja Hofmann, Shimon Whiteson
To rapidly learn a new task, it is often essential for agents to explore efficiently - especially when performance matters from the first timestep. One way to learn such behaviour is via meta-learning. Many existing methods however rely on dense rewards for meta-training, and can fail catastrophically if the rewards are sparse. ■Without a suitable reward signal, the need for exploration during meta-training is exacerbated. To address this, we propose HyperX, which uses novel reward bonuses for meta-training to explore in approximate hyper-state space (where hyper-states represent the environment state and the agent’s task belief). We show empirically that HyperX meta-learns better task-exploration and adapts more successfully to new tasks than existing methods.
****
Provable Robustness of Adversarial Training for Learning Halfspaces with Noise
Difan Zou, Spencer Frei, Quanquan Gu
We analyze the properties of adversarial training for learning adversarially robust halfspaces in the presence of agnostic label noise. Denoting $\mathsf{OPT}_{p,r}$ as the best classification error achieved by a halfspace that is robust to perturbations of ℓ^{p} balls of radius r, we show that adversarial training on the standard binary cross-entropy loss yields adversarially robust halfspaces up to classification error $\tilde O(\sqrt{\mathsf{OPT}_{2,r}})$ for $p=2$, and $\tilde O(d^{1/4} \sqrt{\mathsf{OPT}_{\infty, r}})$ when $p=\infty$. Our results hold for distributions satisfying anti-concentration properties enjoyed by log-concave isotropic distributions among others. We additionally show that if one instead uses a non-convex sigmoidal loss, adversarial training yields halfspaces with an improved robust classification error of $O(\mathsf{OPT}_{2,r})$ for $p=2$, and $O(d^{1/4} \mathsf{OPT}_{\infty, r})$ when $p=\infty$. To the best of our knowledge, this is the first work showing that adversarial training provably yields robust classifiers in the presence of noise.
****
On the Convergence of Hamiltonian Monte Carlo with Stochastic Gradients
Difan Zou, Quanquan Gu
Hamiltonian Monte Carlo (HMC), built based on the Hamilton’s equation, has been witnessed great success in sampling from high-dimensional posterior distributions. However, it also suffers from computational inefficiency, especially for large training datasets. One common idea to overcome this computational bottleneck is using stochastic gradients, which only queries a mini-batch of training data in each iteration. However, unlike the extensive studies on the convergence analysis of HMC using full gradients, few works focus on establishing the convergence guarantees of stochastic gradient HMC algorithms. In this paper, we propose a general framework for proving the convergence rate of HMC with stochastic gradient estimators, for sampling from strongly log-concave and log-smooth target distributions. We show that the convergence to the target distribution in 2-Wasserstein distance can be guaranteed as long as the stochastic gradient estimator is unbiased and its variance is upper bounded along the algorithm trajectory. We further apply the proposed framework to analyze the convergence rates of HMC with four standard stochastic gradient estimators: mini-batch stochastic gradient (SG), stochastic variance reduced gradient (SVRG), stochastic average gradient (SAGA), and control variate gradient (CVG). Theoretical results explain the inefficiency of mini-batch SG, and suggest that SVRG and SAGA perform better in the task

s with high-precision requirements, while CVG performs better for large dataset. Experiment results verify our theoretical findings.
****

A Functional Perspective on Learning Symmetric Functions with Neural Networks
Aaron Zweig, Joan Bruna
Symmetric functions, which take as input an unordered, fixed-size set, are known to be universally representable by neural networks that enforce permutation invariance. These architectures only give guarantees for fixed input sizes, yet in many practical applications, including point clouds and particle physics, a relevant notion of generalization should include varying the input size. In this work we treat symmetric functions (of any size) as functions over probability measures, and study the learning and representation of neural networks defined on measures. By focusing on shallow architectures, we establish approximation and generalization bounds under different choices of regularization (such as RKHS and variation norms), that capture a hierarchy of functional spaces with increasing degree of non-linear learning. The resulting models can be learned efficiently and enjoy generalization guarantees that extend across input sizes, as we verify empirically.
****