Sign in to GitHub · GitHub
****
Improved Regret Bounds for Thompson Sampling in Linear Quadratic Control Problems
Marc Abeille, Alessandro Lazaric
Thompson sampling (TS) is an effective approach to trade off exploration and exploration in reinforcement learning. Despite its empirical success and recent advances, its theoretical analysis is often limited to the Bayesian setting, finite state-action spaces, or finite-horizon problems. In this paper, we study an instance of TS in the challenging setting of the infinite-horizon linear quadratic (LQ) control, which models problems with continuous state-action variables, linear dynamics, and quadratic cost. In particular, we analyze the regret in the frequentist sense (i.e., for a fixed unknown environment) in one-dimensional systems. We derive the first $O(\sqrt{T})$ frequentist regret bound for this problem, thus significantly improving the $O(T^{2/3})$ bound of Abeille & Lazaric (2017) and matching the frequentist performance derived by Abbasi-Yadkori & Szepesvári (2011) for an optimistic approach and the Bayesian result Ouyang et al. (2017) We obtain this result by developing a novel bound on the regret due to policy switches, which holds for LQ systems of any dimensionality and it allows updating the parameters and the policy at each step, thus overcoming previous limitations due to lazy updates. Finally, we report numerical simulations supporting the conjecture that our result extends to multi-dimensional systems.
****
State Abstractions for Lifelong Reinforcement Learning
David Abel, Dilip Arumugam, Lucas Lehnert, Michael Littman
In lifelong reinforcement learning, agents must effectively transfer knowledge across tasks while simultaneously addressing exploration, credit assignment, and generalization. State abstraction can help overcome these hurdles by compressing the representation used by an agent, thereby reducing the computational and statistical burdens of learning. To this end, we here develop theory to compute and use state abstractions in lifelong reinforcement learning. We introduce two new classes of abstractions: (1) transitive state abstractions, whose optimal form can be computed efficiently, and (2) PAC state abstractions, which are guaranteed to hold with respect to a distribution of tasks. We show that the joint family of transitive PAC abstractions can be acquired efficiently, preserve near optimal-behavior, and experimentally reduce sample complexity in simple domains, thereby yielding a family of desirable abstractions for use in lifelong reinforcement learning. Along with these positive results, we show that there are pathological cases where state abstractions can negatively impact performance.
****
Policy and Value Transfer in Lifelong Reinforcement Learning
David Abel, Yuu Jinnai, Sophie Yue Guo, George Konidaris, Michael Littman
We consider the problem of how best to use prior experience to bootstrap lifelong learning, where an agent faces a series of task instances drawn from some task distribution. First, we identify the initial policy that optimizes expected performance over the distribution of tasks for increasingly complex classes of policy and task distributions. We empirically demonstrate the relative performance of each policy class’ optimal element in a variety of simple task distributions. We then consider value-function initialization methods that preserve PAC guarantees while simultaneously minimizing the learning required in two learning algorithms, yielding MaxQInit, a practical new method for value-function-based transfer. We show that MaxQInit performs well in simple lifelong RL experiments.
****
INSPECTRE: Privately Estimating the Unseen
Jayadev Acharya, Gautam Kamath, Ziteng Sun, Huanyu Zhang
We develop differentially private methods for estimating various distributional properties. Given a sample from a discrete distribution p, some functional f, and accuracy and privacy parameters alpha and epsilon, the goal is to estimate f(p) up to accuracy alpha, while maintaining epsilon-differential privacy of the sample. We prove almost-tight bounds on the sample size required for this problem

for several functionals of interest, including support size, support coverage, a
nd entropy. We show that the cost of privacy is negligible in a variety of setti
ngs, both theoretically and experimentally. Our methods are based on a sensitivi
ty analysis of several state-of-the-art methods for estimating these properties
with sublinear sample complexities
****

Learning Representations and Generative Models for 3D Point Clouds
Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, Leonidas Guibas
Three-dimensional geometric data offer an excellent domain for studying represen
tation learning and generative modeling. In this paper, we look at geometric dat
a represented as point clouds. We introduce a deep AutoEncoder (AE) network with
 state-of-the-art reconstruction quality and generalization ability. The learned
 representations outperform existing methods on 3D recognition tasks and enable
shape editing via simple algebraic manipulations, such as semantic part editing,
 shape analogies and shape interpolation, as well as shape completion. We perfor
m a thorough study of different generative models including GANs operating on th
e raw point clouds, significantly improved GANs trained in the fixed latent spac
e of our AEs, and Gaussian Mixture Models (GMMs). To quantitatively evaluate gen
erative models we introduce measures of sample fidelity and diversity based on m
atchings between sets of point clouds. Interestingly, our evaluation of generali
zation, fidelity and diversity reveals that GMMs trained in the latent space of
our AEs yield the best results overall.
****

Discovering Interpretable Representations for Both Deep Generative and Discrimin
ative Models
Tameem Adel, Zoubin Ghahramani, Adrian Weller
Interpretability of representations in both deep generative and discriminative m
odels is highly desirable. Current methods jointly optimize an objective combini
ng accuracy and interpretability. However, this may reduce accuracy, and is not
applicable to already trained models. We propose two interpretability frameworks
. First, we provide an interpretable lens for an existing model. We use a genera
tive model which takes as input the representation in an existing (generative or
 discriminative) model, weakly supervised by limited side information. Applying
a flexible and invertible transformation to the input leads to an interpretable
representation with no loss in accuracy. We extend the approach using an active
learning strategy to choose the most useful side information to obtain, allowing
 a human to guide what "interpretable" means. Our second framework relies on joi
nt optimization for a representation which is both maximally informative about t
he side information and maximally compressive about the non-interpretable data f
actors. This leads to a novel perspective on the relationship between compressio
n and regularization. We also propose a new interpretability evaluation metric b
ased on our framework. Empirically, we achieve state-of-the-art results on three
 datasets using the two proposed algorithms.
****

A Reductions Approach to Fair Classification
Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, Hanna Wallach
We present a systematic approach for achieving fairness in a binary classificati
on setting. While we focus on two well-known quantitative definitions of fairnes
s, our approach encompasses many other previously studied definitions as special
 cases. The key idea is to reduce fair classification to a sequence of cost-sens
itive classification problems, whose solutions yield a randomized classifier wit
h the lowest (empirical) error subject to the desired constraints. We introduce
two reductions that work for any representation of the cost-sensitive classifier
 and compare favorably to prior baselines on a variety of data sets, while overc
oming several of their disadvantages.
****

Accelerated Spectral Ranking
Arpit Agarwal, Prathamesh Patil, Shivani Agarwal
The problem of rank aggregation from pairwise and multiway comparisons has a wid
e range of implications, ranging from recommendation systems to sports rankings

to social choice. Some of the most popular algorithms for this problem come from the class of spectral ranking algorithms; these include the rank centrality (RC) algorithm for pairwise comparisons, which returns consistent estimates under the Bradley-Terry-Luce (BTL) model for pairwise comparisons (Negahban et al., 2017), and its generalization, the Luce spectral ranking (LSR) algorithm, which returns consistent estimates under the more general multinomial logit (MNL) model for multiway comparisons (Maystre & Grossglauser, 2015). In this paper, we design a provably faster spectral ranking algorithm, which we call accelerated spectral ranking (ASR), that is also consistent under the MNL/BTL models. Our accelerated algorithm is achieved by designing a random walk that has a faster mixing time than the random walks associated with previous algorithms. In addition to a faster algorithm, our results yield improved sample complexity bounds for recovery of the MNL/BTL parameters: to the best of our knowledge, we give the first general sample complexity bounds for recovering the parameters of the MNL model from multiway comparisons under any (connected) comparison graph (and improve significantly over previous bounds for the BTL model for pairwise comparisons). We also give a message-passing interpretation of our algorithm, which suggests a decentralized distributed implementation. Our experiments on several real-world and synthetic datasets confirm that our new ASR algorithm is indeed orders of magnitude faster than existing algorithms.
****

MISSION: Ultra Large-Scale Feature Selection using Count-Sketches
Amirali Aghazadeh, Ryan Spring, Daniel Lejeune, Gautam Dasarathy, Anshumali Shrivastava, baraniuk
Feature selection is an important challenge in machine learning. It plays a crucial role in the explainability of machine-driven decisions that are rapidly permeating throughout modern society. Unfortunately, the explosion in the size and dimensionality of real-world datasets poses a severe challenge to standard feature selection algorithms. Today, it is not uncommon for datasets to have billions of dimensions. At such scale, even storing the feature vector is impossible, causing most existing feature selection methods to fail. Workarounds like feature hashing, a standard approach to large-scale machine learning, helps with the computational feasibility, but at the cost of losing the interpretability of features. In this paper, we present MISSION, a novel framework for ultra large-scale feature selection that performs stochastic gradient descent while maintaining an efficient representation of the features in memory using a Count-Sketch data structure. MISSION retains the simplicity of feature hashing without sacrificing the interpretability of the features while using only O(log^2(p)) working memory. We demonstrate that MISSION accurately and efficiently performs feature selection on real-world, large-scale datasets with billions of dimensions.
****

Minimal I-MAP MCMC for Scalable Structure Discovery in Causal DAG Models
Raj Agrawal, Caroline Uhler, Tamara Broderick
Learning a Bayesian network (BN) from data can be useful for decision-making or discovering causal relationships. However, traditional methods often fail in modern applications, which exhibit a larger number of observed variables than data points. The resulting uncertainty about the underlying network as well as the desire to incorporate prior information recommend a Bayesian approach to learning the BN, but the highly combinatorial structure of BNs poses a striking challenge for inference. The current state-of-the-art methods such as order MCMC are faster than previous methods but prevent the use of many natural structural priors and still have running time exponential in the maximum indegree of the true directed acyclic graph (DAG) of the BN. We here propose an alternative posterior approximation based on the observation that, if we incorporate empirical conditional independence tests, we can focus on a high-probability DAG associated with each order of the vertices. We show that our method allows the desired flexibility in prior specification, removes timing dependence on the maximum indegree, and yields provably good posterior approximations; in addition, we show that it achieves superior accuracy, scalability, and sampler mixing on several datasets.
****

Proportional Allocation: Simple, Distributed, and Diverse Matching with High Entropy
Shipra Agrawal, Morteza Zadimoghaddam, Vahab Mirrokni
Inspired by many applications of bipartite matching in online advertising and machine learning, we study a simple and natural iterative proportional allocation algorithm: Maintain a priority score $\priority_a$ for each node $a\in \mathds{A}$ on one side of the bipartition, initialized as $\priority_a=1$. Iteratively allocate the nodes $i\in \impressions$ on the other side to eligible nodes in $\mathds{A}$ in proportion of their priority scores. After each round, for each node $a\in \mathds{A}$, decrease or increase the score $\priority_a$ based on whether it is over- or under- allocated. Our first result is that this simple, distributed algorithm converges to a $(1-\epsilon)$-approximate fractional b-matching solution in $O({\log n\over \epsilon^2})$ rounds. We also extend the proportional allocation algorithm and convergence results to the maximum weighted matching problem, and show that the algorithm can be naturally tuned to produce maximum matching with high entropy. High entropy, in turn, implies additional desirable properties of this matching, e.g., it satisfies certain diversity and fairness (aka anonymity) properties that are desirable in a variety of applications in online advertising and machine learning.
****
Bucket Renormalization for Approximate Inference
Sungsoo Ahn, Michael Chertkov, Adrian Weller, Jinwoo Shin
Probabilistic graphical models are a key tool in machine learning applications. Computing the partition function, i.e., normalizing constant, is a fundamental task of statistical inference but is generally computationally intractable, leading to extensive study of approximation methods. Iterative variational methods are a popular and successful family of approaches. However, even state of the art variational methods can return poor results or fail to converge on difficult instances. In this paper, we instead consider computing the partition function via sequential summation over variables. We develop robust approximate algorithms by combining ideas from mini-bucket elimination with tensor network and renormalization group methods from statistical physics. The resulting “convergence-free” methods show good empirical performance on both synthetic and real-world benchmark models, even for difficult instances.
****
oi-VAE: Output Interpretable VAEs for Nonlinear Group Factor Analysis
Samuel K. Ainsworth, Nicholas J. Foti, Adrian K. C. Lee, Emily B. Fox
Deep generative models have recently yielded encouraging results in producing subjectively realistic samples of complex data. Far less attention has been paid to making these generative models interpretable. In many scenarios, ranging from scientific applications to finance, the observed variables have a natural grouping. It is often of interest to understand systems of interaction amongst these groups, and latent factor models (LFMs) are an attractive approach. However, traditional LFMs are limited by assuming a linear correlation structure. We present an output interpretable VAE (oi-VAE) for grouped data that models complex, nonlinear latent-to-observed relationships. We combine a structured VAE comprised of group-specific generators with a sparsity-inducing prior. We demonstrate that oi-VAE yields meaningful notions of interpretability in the analysis of motion capture and MEG data. We further show that in these situations, the regularization inherent to oi-VAE can actually lead to improved generalization and learned generative processes.
****
Limits of Estimating Heterogeneous Treatment Effects: Guidelines for Practical Algorithm Design
Ahmed Alaa, Mihaela Schaar
Estimating heterogeneous treatment effects from observational data is a central problem in many domains. Because counterfactual data is inaccessible, the problem differs fundamentally from supervised learning, and entails a more complex set of modeling choices. Despite a variety of recently proposed algorithmic solutions, a principled guideline for building estimators of treatment effects using ma

chine learning algorithms is still lacking. In this paper, we provide such a gui deline by characterizing the fundamental limits of estimating heterogeneous trea tment effects, and establishing conditions under which these limits can be achie ved. Our analysis reveals that the relative importance of the different aspects of observational data vary with the sample size. For instance, we show that sele ction bias matters only in small-sample regimes, whereas with a large sample siz e, the way an algorithm models the control and treated outcomes is what bottlene cks its performance. Guided by our analysis, we build a practical algorithm for estimating treatment effects using a non-stationary Gaussian processes with doub ly-robust hyperparameters. Using a standard semi-synthetic simulation setup, we show that our algorithm outperforms the state-of-the-art, and that the behavior of existing algorithms conforms with our analysis.

****

AutoPrognosis: Automated Clinical Prognostic Modeling via Bayesian Optimization with Structured Kernel Learning

Ahmed Alaa, Mihaela Schaar

Clinical prognostic models derived from largescale healthcare data can inform cr itical diagnostic and therapeutic decisions. To enable off-theshelf usage of mac hine learning (ML) in prognostic research, we developed AUTOPROGNOSIS: a system for automating the design of predictive modeling pipelines tailored for clinical prognosis. AUTOPROGNOSIS optimizes ensembles of pipeline configurations efficie ntly using a novel batched Bayesian optimization (BO) algorithm that learns a lo w-dimensional decomposition of the pipelines' high-dimensional hyperparameter sp ace in concurrence with the BO procedure. This is achieved by modeling the pipel ines' performances as a black-box function with a Gaussian process prior, and mo deling the "similarities" between the pipelines' baseline algorithms via a spars e additive kernel with a Dirichlet prior. Meta-learning is used to warmstart BO with external data from "similar" patient cohorts by calibrating the priors usin g an algorithm that mimics the empirical Bayes method. The system automatically explains its predictions by presenting the clinicians with logical association r ules that link patients' features to predicted risk strata. We demonstrate the u tility of AUTOPROGNOSIS using 10 major patient cohorts representing various aspe cts of cardiovascular patient care.

****

Information Theoretic Guarantees for Empirical Risk Minimization with Applicatio ns to Model Selection and Large-Scale Optimization

Ibrahim Alabdulmohsin

In this paper, we derive bounds on the mutual information of the empirical risk minimization (ERM) procedure for both 0-1 and strongly-convex loss classes. We p rove that under the Axiom of Choice, the existence of an ERM learning rule with a vanishing mutual information is equivalent to the assertion that the loss clas s has a finite VC dimension, thus bridging information theory with statistical l earning theory. Similarly, an asymptotic bound on the mutual information is esta blished for strongly-convex loss classes in terms of the number of model paramet ers. The latter result rests on a central limit theorem (CLT) that we derive in this paper. In addition, we use our results to analyze the excess risk in stocha stic convex optimization and unify previous works. Finally, we present two impor tant applications. First, we show that the ERM of strongly-convex loss classes c an be trivially scaled to big data using a naive parallelization algorithm with provable guarantees. Second, we propose a simple information criterion for model selection and demonstrate experimentally that it outperforms the popular Akaike 's information criterion (AIC) and Schwarz's Bayesian information criterion (BIC).

****

Fixing a Broken ELBO

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A. Saurous, Kevin Mu rphy

Recent work in unsupervised representation learning has focused on learning deep directed latentvariable models. Fitting these models by maximizing the marginal likelihood or evidence is typically intractable, thus a common approximation is

to maximize the evidence lower bound (ELBO) instead. However, maximum likelihood training (whether exact or approximate) does not necessarily result in a good latent representation, as we demonstrate both theoretically and empirically. In particular, we derive variational lower and upper bounds on the mutual information between the input and the latent variable, and use these bounds to derive a rate-distortion curve that characterizes the tradeoff between compression and reconstruction accuracy. Using this framework, we demonstrate that there is a family of models with identical ELBO, but different quantitative and qualitative characteristics. Our framework also suggests a simple new method to ensure that latent variable models with powerful stochastic decoders do not ignore their latent code.

****

Differentially Private Identity and Equivalence Testing of Discrete Distributions

Maryam Aliakbarpour, Ilias Diakonikolas, Ronitt Rubinfeld

We study the fundamental problems of identity and equivalence testing over a discrete population from random samples. Our goal is to develop efficient testers while guaranteeing differential privacy to the individuals of the population. We provide sample-efficient differentially private testers for these problems. Our theoretical results significantly improve over the best known algorithms for identity testing, and are the first results for private equivalence testing. The conceptual message of our work is that there exist private hypothesis testers that are nearly as sample-efficient as their non-private counterparts. We perform an experimental evaluation of our algorithms on synthetic data. Our experiments illustrate that our private testers achieve small type I and type II errors with sample size sublinear in the domain size of the underlying distributions.

****

Katyusha X: Simple Momentum Method for Stochastic Sum-of-Nonconvex Optimization

Zeyuan Allen-Zhu

The problem of minimizing sum-of-nonconvex functions (i.e., convex functions that are average of non-convex ones) is becoming increasing important in machine learning, and is the core machinery for PCA, SVD, regularized Newton’s method, accelerated non-convex optimization, and more. We show how to provably obtain an accelerated stochastic algorithm for minimizing sum-of-nonconvex functions, by adding one additional line to the well-known SVRG method. This line corresponds to momentum, and shows how to directly apply momentum to the finite-sum stochastic minimization of sum-of-nonconvex functions. As a side result, our method enjoys linear parallel speed-up using mini-batch.

****

Make the Minority Great Again: First-Order Regret Bound for Contextual Bandits

Zeyuan Allen-Zhu, Sebastien Bubeck, Yuanzhi Li

Regret bounds in online learning compare the player’s performance to $L*$, the optimal performance in hindsight with a fixed strategy. Typically such bounds scale with the square root of the time horizon T. The more refined concept of first-order regret bound replaces this with a scaling $\sqrt{L*}$, which may be much smaller than $\sqrt{T}$. It is well known that minor variants of standard algorithms satisfy first-order regret bounds in the full information and multi-armed bandit settings. In a COLT 2017 open problem, Agarwal, Krishnamurthy, Langford, Luo, and Schapire raised the issue that existing techniques do not seem sufficient to obtain first-order regret bounds for the contextual bandit problem. In the present paper, we resolve this open problem by presenting a new strategy based on augmenting the policy space.

****

Augmented CycleGAN: Learning Many-to-Many Mappings from Unpaired Data

Amjad Almahairi, Sai Rajeshwar, Alessandro Sordoni, Philip Bachman, Aaron Courville

Learning inter-domain mappings from unpaired data can improve performance in structured prediction tasks, such as image segmentation, by reducing the need for paired data. CycleGAN was recently proposed for this problem, but critically assumes the underlying inter-domain mapping is approximately deterministic and one-t

o-one. This assumption renders the model ineffective for tasks requiring flexibl e, many-to-many mappings. We propose a new model, called Augmented CycleGAN, whi ch learns many-to-many mappings between domains. We examine Augmented CycleGAN q ualitatively and quantitatively on several image datasets.
****
Meta-Learning by Adjusting Priors Based on Extended PAC-Bayes Theory
Ron Amit, Ron Meir
In meta-learning an agent extracts knowledge from observed tasks, aiming to faci litate learning of novel future tasks. Under the assumption that future tasks ar e ‘related’ to previous tasks, accumulated knowledge should be learned in such a way that they capture the common structure across learned tasks, while allowing the learner sufficient flexibility to adapt to novel aspects of a new task. We present a framework for meta-learning that is based on generalization error boun ds, allowing us to extend various PAC-Bayes bounds to meta-learning. Learning ta kes place through the construction of a distribution over hypotheses based on th e observed tasks, and its utilization for learning a new task. Thus, prior knowl edge is incorporated through setting an experience-dependent prior for novel tas ks. We develop a gradient-based algorithm, and implement it for deep neural netw orks, based on minimizing an objective function derived from the bounds, and dem onstrate its effectiveness numerically. In addition to establishing the improved performance available through meta-learning, we demonstrate the intuitive way b y which prior information is manifested at different levels of the network.
****
MAGAN: Aligning Biological Manifolds
Matthew Amodio, Smita Krishnaswamy
It is increasingly common in many types of natural and physical systems (especia lly biological systems) to have different types of measurements performed on the same underlying system. In such settings, it is important to align the manifold s arising from each measurement in order to integrate such data and gain an impr oved picture of the system; we tackle this problem using generative adversarial networks (GANs). Recent attempts to use GANs to find correspondences between set s of samples do not explicitly perform proper alignment of manifolds. We present the new Manifold Aligning GAN (MAGAN) that aligns two manifolds such that relat ed points in each measurement space are aligned. We demonstrate applications of MAGAN in single-cell biology in integrating two different measurement types toge ther: cells from the same tissue are measured with both genomic (single-cell RNA -sequencing) and proteomic (mass cytometry) technologies. We show that MAGAN suc cessfully aligns manifolds such that known correlations between measured markers are improved compared to other recently proposed models.
****
Subspace Embedding and Linear Regression with Orlicz Norm
Alexandr Andoni, Chengyu Lin, Ying Sheng, Peilin Zhong, Ruiqi Zhong
We consider a generalization of the classic linear regression problem to the cas e when the loss is an Orlicz norm. An Orlicz norm is parameterized by a non-nega tive convex function G: R_+ - > R_+ with G(0) = 0: the Orlicz norm of a n-dimens ional vector x is defined as |x|_G = inf{ alpha > 0 | sum_{i = 1}^n G(|x_i| / a lpha) < = 1 }. We consider the cases where the function G grows subquadraticall y. Our main result is based on a new oblivious embedding which embeds the column space of a given nxd matrix A with Orlicz norm into a lower dimensional space w ith L2 norm. Specifically, we show how to efficiently find an mxn embedding matr ix S (m < n), such that for every d-dimensional vector x, we have Omega(1/(d log n)) |Ax|_G < = |SAx|_2 < = O(d^2 log n) |Ax|_G. By applying this subspace embed ding technique, we show an approximation algorithm for the regression problem mi n_x |Ax-b|_G, up to a O(d log^2 n) factor. As a further application of our tec hniques, we show how to also use them to improve on the algorithm for the Lp low rank matrix approximation problem for 1 < = p < 2.
****
Efficient Gradient-Free Variational Inference using Policy Search
Oleg Arenz, Gerhard Neumann, Mingjun Zhong
Inference from complex distributions is a common problem in machine learning nee

ded for many Bayesian methods. We propose an efficient, gradient-free method for learning general GMM approximations of multimodal distributions based on recent insights from stochastic search methods. Our method establishes information-geometric trust regions to ensure efficient exploration of the sampling space and stability of the GMM updates, allowing for efficient estimation of multi-variate Gaussian variational distributions. For GMMs, we apply a variational lower bound to decompose the learning objective into sub-problems given by learning the individual mixture components and the coefficients. The number of mixture components is adapted online in order to allow for arbitrary exact approximations. We demonstrate on several domains that we can learn significantly better approximations than competing variational inference methods and that the quality of samples drawn from our approximations is on par with samples created by state-of-the-art MCMC samplers that require significantly more computational resources.

****

On the Optimization of Deep Networks: Implicit Acceleration by Overparameterization

Sanjeev Arora, Nadav Cohen, Elad Hazan

Conventional wisdom in deep learning states that increasing depth improves expressiveness but complicates optimization. This paper suggests that, sometimes, increasing depth can speed up optimization. The effect of depth on optimization is decoupled from expressiveness by focusing on settings where additional layers amount to overparameterization – linear neural networks, a well-studied model. Theoretical analysis, as well as experiments, show that here depth acts as a preconditioner which may accelerate convergence. Even on simple convex problems such as linear regression with ℓ_p loss, $p>2$, gradient descent can benefit from transitioning to a non-convex overparameterized objective, more than it would from some common acceleration schemes. We also prove that it is mathematically impossible to obtain the acceleration effect of overparametrization via gradients of any regularizer.

****

Stronger Generalization Bounds for Deep Nets via a Compression Approach

Sanjeev Arora, Rong Ge, Behnam Neyshabur, Yi Zhang

Deep nets generalize well despite having more parameters than the number of training samples. Recent works try to give an explanation using PAC-Bayes and Margin-based analyses, but do not as yet result in sample complexity bounds better than naive parameter counting. The current paper shows generalization bounds that are orders of magnitude better in practice. These rely upon new succinct reparametrizations of the trained net — a compression that is explicit and efficient. These yield generalization bounds via a simple compression-based framework introduced here. Our results also provide some theoretical justification for widespread empirical success in compressing deep nets. Analysis of correctness of our compression relies upon some newly identified noise stability properties of trained deep nets, which are also experimentally verified. The study of these properties and resulting generalization bounds are also extended to convolutional nets, which had eluded earlier attempts on proving generalization.

****

Lipschitz Continuity in Model-based Reinforcement Learning

Kavosh Asadi, Dipendra Misra, Michael Littman

We examine the impact of learning Lipschitz continuous models in the context of model-based reinforcement learning. We provide a novel bound on multi-step prediction error of Lipschitz models where we quantify the error using the Wasserstein metric. We go on to prove an error bound for the value-function estimate arising from Lipschitz models and show that the estimated value function is itself Lipschitz. We conclude with empirical results that show the benefits of controlling the Lipschitz constant of neural-network models.

****

Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

Anish Athalye, Nicholas Carlini, David Wagner

We identify obfuscated gradients, a kind of gradient masking, as a phenomenon th

at leads to a false sense of security in defenses against adversarial examples. While defenses that cause obfuscated gradients appear to defeat iterative optimi zation-based attacks, we find defenses relying on this effect can be circumvente d. We describe characteristic behaviors of defenses exhibiting the effect, and f or each of the three types of obfuscated gradients we discover, we develop attac k techniques to overcome it. In a case study, examining non-certified white-box- secure defenses at ICLR 2018, we find obfuscated gradients are a common occurren ce, with 7 of 9 defenses relying on obfuscated gradients. Our new attacks succes sfully circumvent 6 completely, and 1 partially, in the original threat model ea ch paper considers.

****

Synthesizing Robust Adversarial Examples

Anish Athalye, Logan Engstrom, Andrew Ilyas, Kevin Kwok

Standard methods for generating adversarial examples for neural networks do not consistently fool neural network classifiers in the physical world due to a comb ination of viewpoint shifts, camera noise, and other natural transformations, li miting their relevance to real-world systems. We demonstrate the existence of ro bust 3D adversarial objects, and we present the first algorithm for synthesizing examples that are adversarial over a chosen distribution of transformations. We synthesize two-dimensional adversarial images that are robust to noise, distort ion, and affine transformation. We apply our algorithm to complex three-dimensio nal objects, using 3D-printing to manufacture the first physical adversarial obj ects. Our results demonstrate the existence of 3D adversarial objects in the phy sical world.

****

Contextual Graph Markov Model: A Deep and Generative Approach to Graph Processin g

Davide Bacciu, Federico Errica, Alessio Micheli

We introduce the Contextual Graph Markov Model, an approach combining ideas from generative models and neural networks for the processing of graph data. It foun ds on a constructive methodology to build a deep architecture comprising layers of probabilistic models that learn to encode the structured information in an in cremental fashion. Context is diffused in an efficient and scalable way across t he graph vertexes and edges. The resulting graph encoding is used in combination with discriminative models to address structure classification benchmarks.

****

Greed is Still Good: Maximizing Monotone Submodular+Supermodular (BP) Functions

Wenruo Bai, Jeff Bilmes

We analyze the performance of the greedy algorithm, and also a discrete semi-gra dient based algorithm, for maximizing the sum of a suBmodular and suPermodular (BP) function (both of which are non-negative monotone non-decreasing) under two types of constraints, either a cardinality constraint or $p\geq 1$ matroid indep endence constraints. These problems occur naturally in several real-world applic ations in data science, machine learning, and artificial intelligence. The probl ems are ordinarily inapproximable to any factor. Using the curvature $\curv_f$ o f the submodular term, and introducing $\curv^g$ for the supermodular term (a na tural dual curvature for supermodular functions), however, both of which are com putable in linear time, we show that BP maximization can be efficiently approxim ated by both the greedy and the semi-gradient based algorithm. The algorithms yi eld multiplicative guarantees of $\frac{1}{\curv_f}\left[1-e^{-(1-\curv^g)\curv_ f}\right]$ and $\frac{1-\curv^g}{(1-\curv^g)\curv_f + p}$ for the two types of c onstraints respectively. For pure monotone supermodular constrained maximization , these yield $1-\curvg$ and $(1-\curvg)/p$ for the two types of constraints res pectively. We also analyze the hardness of BP maximization and show that our gua rantees match hardness by a constant factor and by $O(\ln(p))$ respectively. Com putational experiments are also provided supporting our analysis.

****

Comparing Dynamics: Deep Neural Networks versus Glassy Systems

Marco Baity-Jesi, Levent Sagun, Mario Geiger, Stefano Spigler, Gerard Ben Arous, Chiara Cammarota, Yann LeCun, Matthieu Wyart, Giulio Biroli

We analyze numerically the training dynamics of deep neural networks (DNN) by us ing methods developed in statistical physics of glassy systems. The two main iss ues we address are the complexity of the loss-landscape and of the dynamics with in it, and to what extent DNNs share similarities with glassy systems. Our findi ngs, obtained for different architectures and data-sets, suggest that during the training process the dynamics slows down because of an increasingly large numbe r of flat directions. At large times, when the loss is approaching zero, the sys tem diffuses at the bottom of the landscape. Despite some similarities with the dynamics of mean-field glassy systems, in particular, the absence of barrier cro ssing, we find distinctive dynamical behaviors in the two cases, thus showing th at the statistical properties of the corresponding loss and energy landscapes ar e different. In contrast, when the network is under-parametrized we observe a ty pical glassy behavior, thus suggesting the existence of different phases dependi ng on whether the network is under-parametrized or over-parametrized.

****

SMAC: Simultaneous Mapping and Clustering Using Spectral Decompositions

Chandrajit Bajaj, Tingran Gao, Zihang He, Qixing Huang, Zhenxiao Liang

We introduce a principled approach for simultaneous mapping and clustering (SMAC) for establishing consistent maps across heterogeneous object collections (e.g. , 2D images or 3D shapes). Our approach takes as input a heterogeneous object co llection and a set of maps computed between some pairs of objects, and outputs a homogeneous object clustering together with a new set of maps possessing optima l intra- and inter-cluster consistency. Our approach is based on the spectral de composition of a data matrix storing all pairwise maps in its blocks. We additio nally provide tight theoretical guarantees on the exactness of SMAC under establ ished noise models. We also demonstrate the usefulness of the approach on synthe tic and real datasets.

****

A Boo(n) for Evaluating Architecture Performance

Ondrej Bajgar, Rudolf Kadlec, Jan Kleindienst

We point out important problems with the common practice of using the best singl e model performance for comparing deep learning architectures, and we propose a method that corrects these flaws. Each time a model is trained, one gets a diffe rent result due to random factors in the training process, which include random parameter initialization and random data shuffling. Reporting the best single mo del performance does not appropriately address this stochasticity. We propose a normalized expected best-out-of-n performance (Boo_n) as a way to cor rect these problems.

****

Learning to Branch

Maria-Florina Balcan, Travis Dick, Tuomas Sandholm, Ellen Vitercik

Tree search algorithms, such as branch-and-bound, are the most widely used tools for solving combinatorial problems. These algorithms recursively partition the search space to find an optimal solution. To keep the tree small, it is crucial to carefully decide, when expanding a tree node, which variable to branch on at that node to partition the remaining space. Many partitioning techniques have be en proposed, but no theory describes which is optimal. We show how to use machin e learning to determine an optimal weighting of any set of partitioning procedur es for the instance distribution at hand using samples. Via theory and experimen ts, we show that learning to branch is both practical and hugely beneficial.

****

The Mechanics of n-Player Differentiable Games

David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, Thore Graepel

The cornerstone underpinning deep learning is the guarantee that gradient descen t on an objective converges to local minima. Unfortunately, this guarantee fails in settings, such as generative adversarial nets, where there are multiple inte racting losses. The behavior of gradient-based methods in games is not well unde rstood – and is becoming increasingly important as adversarial and multi-objecti ve architectures proliferate. In this paper, we develop new techniques to unders

tand and control the dynamics in general games. The key result is to decompose t
he second-order dynamics into two components. The first is related to potential
games, which reduce to gradient descent on an implicit function; the second rela
tes to Hamiltonian games, a new class of games that obey a conservation law, aki
n to conservation laws in classical mechanical systems. The decomposition motiva
tes Symplectic Gradient Adjustment (SGA), a new algorithm for finding stable fix
ed points in general games. Basic experiments show SGA is competitive with recen
tly proposed algorithms for finding local Nash equilibria in GANs – whilst at th
e same time being applicable to – and having guarantees in – much more general g
ames.
****

Spline Filters For End-to-End Deep Learning
Randall Balestriero, Romain Cosentino, Herve Glotin, Richard Baraniuk
We propose to tackle the problem of end-to-end learning for raw waveform signals
 by introducing learnable continuous time-frequency atoms. The derivation of the
se filters is achieved by defining a functional space with a given smoothness or
der and boundary conditions. From this space, we derive the parametric analytica
l filters. Their differentiability property allows gradient-based optimization.
As such, one can utilize any Deep Neural Network (DNN) with these filters. This
enables us to tackle in a front-end fashion a large scale bird detection task ba
sed on the freefield1010 dataset known to contain key challenges, such as the di
mensionality of the inputs data ($>100,000$) and the presence of additional nois
es: multiple sources and soundscapes.
****

A Spline Theory of Deep Learning
Randall Balestriero, baraniuk
We build a rigorous bridge between deep networks (DNs) and approximation theory
via spline functions and operators. Our key result is that a large class of DNs
can be written as a composition of max-affine spline operators (MASOs), which pr
ovide a powerful portal through which to view and analyze their inner workings.
For instance, conditioned on the input signal, the output of a MASO DN can be wr
itten as a simple affine transformation of the input. This implies that a DN con
structs a set of signal-dependent, class-specific templates against which the si
gnal is compared via a simple inner product; we explore the links to the classic
al theory of optimal classification via matched filters and the effects of data
memorization. Going further, we propose a simple penalty term that can be added
to the cost function of any DN learning algorithm to force the templates to be o
rthogonal with each other; this leads to significantly improved classification p
erformance and reduced overfitting with no change to the DN architecture. The sp
line partition of the input signal space opens up a new geometric avenue to stud
y how DNs organize signals in a hierarchical fashion. As an application, we deve
lop and validate a new distance metric for signals that quantifies the differenc
e between their partition encodings.
****

Approximation Guarantees for Adaptive Sampling
Eric Balkanski, Yaron Singer
In this paper we analyze an adaptive sampling approach for submodular maximizati
on. Adaptive sampling is a technique that has recently been shown to achieve a c
onstant factor approximation guarantee for submodular maximization under a cardi
nality constraint with exponentially fewer adaptive rounds than any previously s
tudied constant factor approximation algorithm for this problem. Adaptivity quan
tifies the number of sequential rounds that an algorithm makes when function eva
luations can be executed in parallel and is the parallel running time of an algo
rithm, up to low order terms. Adaptive sampling achieves its exponential speedup
 at the expense of approximation. In theory, it is guaranteed to produce a solut
ion that is a 1/3 approximation to the optimum. Nevertheless, experiments show t
hat adaptive sampling techniques achieve far better values in practice. In this
paper we provide theoretical justification for this phenomenon. In particular, w
e show that under very mild conditions of curvature of a function, adaptive samp
ling techniques achieve an approximation arbitrarily close to 1/2 while maintain

ing their low adaptivity. Furthermore, we show that the approximation ratio appr
oaches 1 in direct relationship to a homogeneity property of the submodular func
tion. In addition, we conduct experiments on real data sets in which the curvatu
re and homogeneity properties can be easily manipulated and demonstrate the rela
tionship between approximation and curvature, as well as the effectiveness of ad
aptive sampling in practice.
****

Improving the Gaussian Mechanism for Differential Privacy: Analytical Calibratio
n and Optimal Denoising
Borja Balle, Yu-Xiang Wang
The Gaussian mechanism is an essential building block used in multitude of diffe
rentially private data analysis algorithms. In this paper we revisit the Gaussia
n mechanism and show that the original analysis has several important limitation
s. Our analysis reveals that the variance formula for the original mechanism is
far from tight in the high privacy regime ($\varepsilon \to 0$) and it cannot be
 extended to the low privacy regime ($\varepsilon \to \infty$). We address these
 limitations by developing an optimal Gaussian mechanism whose variance is calib
rated directly using the Gaussian cumulative density function instead of a tail
bound approximation. We also propose to equip the Gaussian mechanism with a post
-processing step based on adaptive estimation techniques by leveraging that the
distribution of the perturbation is known. Our experiments show that analytical
calibration removes at least a third of the variance of the noise compared to th
e classical Gaussian mechanism, and that denoising dramatically improves the acc
uracy of the Gaussian mechanism in the high-dimensional regime.
****

Dissecting Adam: The Sign, Magnitude and Variance of Stochastic Gradients
Lukas Balles, Philipp Hennig
The ADAM optimizer is exceedingly popular in the deep learning community. Often
it works very well, sometimes it doesn’t. Why? We interpret ADAM as a combinatio
n of two aspects: for each weight, the update direction is determined by the sig
n of stochastic gradients, whereas the update magnitude is determined by an esti
mate of their relative variance. We disentangle these two aspects and analyze th
em in isolation, gaining insight into the mechanisms underlying ADAM. This analy
sis also extends recent results on adverse effects of ADAM on generalization, is
olating the sign aspect as the problematic one. Transferring the variance adapta
tion to SGD gives rise to a novel method, completing the practitioner’s toolbox
for problems where ADAM fails.
****

Differentially Private Database Release via Kernel Mean Embeddings
Matej Balog, Ilya Tolstikhin, Bernhard Schölkopf
We lay theoretical foundations for new database release mechanisms that allow th
ird-parties to construct consistent estimators of population statistics, while e
nsuring that the privacy of each individual contributing to the database is prot
ected. The proposed framework rests on two main ideas. First, releasing (an esti
mate of) the kernel mean embedding of the data generating random variable instea
d of the database itself still allows third-parties to construct consistent esti
mators of a wide class of population statistics. Second, the algorithm can satis
fy the definition of differential privacy by basing the released kernel mean emb
edding on entirely synthetic data points, while controlling accuracy through the
 metric available in a Reproducing Kernel Hilbert Space. We describe two instant
iations of the proposed framework, suitable under different scenarios, and prove
 theoretical results guaranteeing differential privacy of the resulting algorith
ms and the consistency of estimators constructed from their outputs.
****

Improving Optimization for Models With Continuous Symmetry Breaking
Robert Bamler, Stephan Mandt
Many loss functions in representation learning are invariant under a continuous
symmetry transformation. For example, the loss function of word embeddings (Miko
lov et al., 2013) remains unchanged if we simultaneously rotate all word and con
text embedding vectors. We show that representation learning models for time ser

ies possess an approximate continuous symmetry that leads to slow convergence of gradient descent. We propose a new optimization algorithm that speeds up convergence using ideas from gauge theory in physics. Our algorithm leads to orders of magnitude faster convergence and to more interpretable representations, as we show for dynamic extensions of matrix factorization and word embedding models. We further present an example application of our proposed algorithm that translates modern words into their historic equivalents.
****
Improved Training of Generative Adversarial Networks Using Representative Features
Duhyeon Bang, Hyunjung Shim
Despite the success of generative adversarial networks (GANs) for image generation, the trade-off between visual quality and image diversity remains a significant issue. This paper achieves both aims simultaneously by improving the stability of training GANs. The key idea of the proposed approach is to implicitly regularize the discriminator using representative features. Focusing on the fact that standard GAN minimizes reverse Kullback-Leibler (KL) divergence, we transfer the representative feature, which is extracted from the data distribution using a pre-trained autoencoder (AE), to the discriminator of standard GANs. Because the AE learns to minimize forward KL divergence, our GAN training with representative features is influenced by both reverse and forward KL divergence. Consequently, the proposed approach is verified to improve visual quality and diversity of state of the art GANs using extensive evaluations.
****
Using Inherent Structures to design Lean 2-layer RBMs
Abhishek Bansal, Abhinav Anand, Chiranjib Bhattacharyya
Understanding the representational power of Restricted Boltzmann Machines (RBMs) with multiple layers is an ill-understood problem and is an area of active research. Motivated from the approach of Inherent Structure formalism (Stillinger & Weber, 1982), extensively used in analysing Spin Glasses, we propose a novel measure called Inherent Structure Capacity (ISC), which characterizes the representation capacity of a fixed architecture RBM by the expected number of modes of distributions emanating from the RBM with parameters drawn from a prior distribution. Though ISC is intractable, we show that for a single layer RBM architecture ISC approaches a finite constant as number of hidden units are increased and to further improve the ISC, one needs to add a second layer. Furthermore, we introduce Lean RBMs, which are multi-layer RBMs where each layer can have at-most O(n) units with the number of visible units being n. We show that for every single layer RBM with Omega(n^{2+r}), r >= 0, hidden units there exists a two-layered lean RBM with Theta(n^2) parameters with the same ISC, establishing that 2 layer RBMs can achieve the same representational power as single-layer RBMs but using far fewer number of parameters. To the best of our knowledge, this is the first result which quantitatively establishes the need for layering.
****
Classification from Pairwise Similarity and Unlabeled Data
Han Bao, Gang Niu, Masashi Sugiyama
Supervised learning needs a huge amount of labeled data, which can be a big bottleneck under the situation where there is a privacy concern or labeling cost is high. To overcome this problem, we propose a new weakly-supervised learning setting where only similar (S) data pairs (two examples belong to the same class) and unlabeled (U) data points are needed instead of fully labeled data, which is called SU classification. We show that an unbiased estimator of the classification risk can be obtained only from SU data, and the estimation error of its empirical risk minimizer achieves the optimal parametric convergence rate. Finally, we demonstrate the effectiveness of the proposed method through experiments.
****
Bayesian Optimization of Combinatorial Structures
Ricardo Baptista, Matthias Poloczek
The optimization of expensive-to-evaluate black-box functions over combinatorial structures is an ubiquitous task in machine learning, engineering and the natur

al sciences. The combinatorial explosion of the search space and costly evaluati ons pose challenges for current techniques in discrete optimization and machine learning, and critically require new algorithmic ideas. This article proposes, t o the best of our knowledge, the first algorithm to overcome these challenges, b ased on an adaptive, scalable model that identifies useful combinatorial structu re even when data is scarce. Our acquisition function pioneers the use of semide finite programming to achieve efficiency and scalability. Experimental evaluatio ns demonstrate that this algorithm consistently outperforms other methods from c ombinatorial and Bayesian optimization.
****

Geodesic Convolutional Shape Optimization
Pierre Baque, Edoardo Remelli, Francois Fleuret, Pascal Fua
Aerodynamic shape optimization has many industrial applications. Existing method s, however, are so computationally demanding that typical engineering practices are to either simply try a limited number of hand-designed shapes or restrict on eself to shapes that can be parameterized using only few degrees of freedom. In this work, we introduce a new way to optimize complex shapes fast and accurately . To this end, we train Geodesic Convolutional Neural Networks to emulate a flui dynamics simulator. The key to making this approach practical is remeshing the o riginal shape using a poly-cube map, which makes it possible to perform the comp utations on GPUs instead of CPUs. The neural net is then used to formulate an ob jective function that is differentiable with respect to the shape parameters, wh ich can then be optimized using a gradient-based technique. This outperforms sta te-of-the-art methods by 5 to 20% for standard problems and, even more important ly, our approach applies to cases that previous methods cannot handle.
****

Learning to Coordinate with Coordination Graphs in Repeated Single-Stage Multi-A gent Decision Problems
Eugenio Bargiacchi, Timothy Verstraeten, Diederik Roijers, Ann Nowé, Hado Hassel t
Learning to coordinate between multiple agents is an important problem in many r einforcement learning problems. Key to learning to coordinate is exploiting loos e couplings, i.e., conditional independences between agents. In this paper we st udy learning in repeated fully cooperative games, multi-agent multi-armed bandit s (MAMABs), in which the expected rewards can be expressed as a coordination gra ph. We propose multi-agent upper confidence exploration (MAUCE), a new algorithm for MAMABs that exploits loose couplings, which enables us to prove a regret bo und that is logarithmic in the number of arm pulls and only linear in the number of agents. We empirically compare MAUCE to sparse cooperative Q-learning, and a state-of-the-art combinatorial bandit approach, and show that it performs much better on a variety of settings, including learning control policies for wind fa rms.
****

Testing Sparsity over Known and Unknown Bases
Siddharth Barman, Arnab Bhattacharyya, Suprovat Ghoshal
Sparsity is a basic property of real vectors that is exploited in a wide variety of machine learning applications. In this work, we describe property testing al gorithms for sparsity that observe a low-dimensional projec- tion of the input. We consider two settings. In the first setting, we test sparsity with respect to an unknown basis: given input vectors $y_1 ,...,y_p \in R^d$ whose concatenatio n as columns forms $Y \in R^{d \times p}$, does $Y = AX$ for matrices $A \in R^ {d\times m}$ and $X \in R^{m \times p}$ such that each column of X is k-spar se, or is Y “far” from having such a decomposition? In the second setting, we test sparsity with respect to a known basis: for a fixed design ma- trix $A \in R^{d \times m}$, given input vector $y \in R^d$, is $y = Ax$ for some k-spar se vector x or is y “far” from having such a decomposition? We analyze our a lgorithms using tools from high-dimensional geometry and probability.
****

Transfer in Deep Reinforcement Learning Using Successor Features and Generalised Policy Improvement

Andre Barreto, Diana Borsa, John Quan, Tom Schaul, David Silver, Matteo Hessel, Daniel Mankowitz, Augustin Zidek, Remi Munos

The ability to transfer skills across tasks has the potential to scale up reinforcement learning (RL) agents to environments currently out of reach. Recently, a framework based on two ideas, successor features (SFs) and generalised policy improvement (GPI), has been introduced as a principled way of transferring skills. In this paper we extend the SF&GPI framework in two ways. One of the basic assumptions underlying the original formulation of SF&GPI is that rewards for all tasks of interest can be computed as linear combinations of a fixed set of features. We relax this constraint and show that the theoretical guarantees supporting the framework can be extended to any set of tasks that only differ in the reward function. Our second contribution is to show that one can use the reward functions themselves as features for future tasks, without any loss of expressiveness, thus removing the need to specify a set of features beforehand. This makes it possible to combine SF&GPI with deep learning in a more stable way. We empirically verify this claim on a complex 3D environment where observations are images from a first-person perspective. We show that the transfer promoted by SF&GPI leads to very good policies on unseen tasks almost instantaneously. We also describe how to learn policies specialised to the new tasks in a way that allows them to be added to the agent's set of skills, and thus be reused in the future.

****

Measuring abstract reasoning in neural networks

David Barrett, Felix Hill, Adam Santoro, Ari Morcos, Timothy Lillicrap

Whether neural networks can learn abstract reasoning or whether they merely rely on superficial statistics is a topic of recent debate. Here, we propose a dataset and challenge designed to probe abstract reasoning, inspired by a well-known human IQ test. To succeed at this challenge, models must cope with various generalisation 'regimes' in which the training data and test questions differ in clearly-defined ways. We show that popular models such as ResNets perform poorly, even when the training and test sets differ only minimally, and we present a novel architecture, with structure designed to encourage reasoning, that does significantly better. When we vary the way in which the test questions and training data differ, we find that our model is notably proficient at certain forms of generalisation, but notably weak at others. We further show that the model's ability to generalise improves markedly if it is trained to predict symbolic explanations for its answers. Altogether, we introduce and explore ways to both measure and induce stronger abstract reasoning in neural networks. Our freely-available dataset should motivate further progress in this direction.

****

Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks

Peter Bartlett, Dave Helmbold, Philip Long

We analyze algorithms for approximating a function $f(x) = \Phi x$ mapping $\Re^d$ to $\Re^d$ using deep linear neural networks, i.e. that learn a function h parameterized by matrices $\Theta_1,...,\Theta_L$ and defined by $h(x) = \Theta_L \Theta_{L-1} ... \Theta_1 x$. We focus on algorithms that learn through gradient descent on the population quadratic loss in the case that the distribution over the inputs is isotropic. We provide polynomial bounds on the number of iterations for gradient descent to approximate the least squares matrix Φ, in the case where the initial hypothesis $\Theta_1 = ... = \Theta_L = I$ has excess loss bounded by a small enough constant. On the other hand, we show that gradient descent fails to converge for Φ whose distance from the identity is a larger constant, and we show that some forms of regularization toward the identity in each layer do not help. If Φ is symmetric positive definite, we show that an algorithm that initializes $\Theta_i = I$ learns an ϵ-approximation of f using a number of updates polynomial in L, the condition number of Φ, and $\log(d/\epsilon)$. In contrast, we show that if the least squares matrix Φ is symmetric and has a negative eigenvalue, then all members of a class of algorithms that perform gradient descent with identity initialization, and optionally regularize toward the identity in each layer, fail to converge. We ana

lyze an algorithm for the case that Φ satisfies $u^{\top} \Phi u > 0$ for a ll u, but may not be symmetric. This algorithm uses two regularizers: one that maintains the invariant $u^{\top} \Theta_L \Theta_{L-1} ... \Theta_1 u > 0$ for all u, and another that "balances" $\Theta_1, ..., \Theta_L$ so that they hav e the same singular values.
****

Mutual Information Neural Estimation
Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, Devon Hjelm
We argue that the estimation of mutual information between high dimensional cont inuous random variables can be achieved by gradient descent over neural networks . We present a Mutual Information Neural Estimator (MINE) that is linearly scala ble in dimensionality as well as in sample size, trainable through back-prop, an d strongly consistent. We present a handful of applications on which MINE can be used to minimize or maximize mutual information. We apply MINE to improve adver sarially trained generative models. We also use MINE to implement the Informatio n Bottleneck, applying it to supervised classification; our results demonstrate substantial improvement in flexibility and performance in these settings.
****

To Understand Deep Learning We Need to Understand Kernel Learning
Mikhail Belkin, Siyuan Ma, Soumik Mandal
Generalization performance of classifiers in deep learning has recently become a subject of intense study. Deep models, which are typically heavily over-paramet rized, tend to fit the training data exactly. Despite this “overfitting", they p erform well on test data, a phenomenon not yet fully understood. The first point of our paper is that strong performance of overfitted classifiers is not a uniq ue feature of deep learning. Using six real-world and two synthetic datasets, we establish experimentally that kernel machines trained to have zero classificati on error or near zero regression error (interpolation) perform very well on test data. We proceed to give a lower bound on the norm of zero loss solutions for s mooth kernels, showing that they increase nearly exponentially with data size. N one of the existing bounds produce non-trivial results for interpolating solutio ns. We also show experimentally that (non-smooth) Laplacian kernels easily fit r andom labels, a finding that parallels results recently reported for ReLU neural networks. In contrast, fitting noisy data requires many more epochs for smooth Gaussian kernels. Similar performance of overfitted Laplacian and Gaussian class ifiers on test, suggests that generalization is tied to the properties of the ke rnel function rather than the optimization process. Some key phenomena of deep l earning are manifested similarly in kernel methods in the modern “overfitted" re gime. The combination of the experimental and theoretical results presented in t his paper indicates a need for new theoretical ideas for understanding propertie s of classical kernel methods. We argue that progress on understanding deep lear ning will be difficult until more tractable “shallow” kernel methods are better understood.
****

Understanding and Simplifying One-Shot Architecture Search
Gabriel Bender, Pieter-Jan Kindermans, Barret Zoph, Vijay Vasudevan, Quoc Le
There is growing interest in automating neural network architecture design. Exis ting architecture search methods can be computationally expensive, requiring tho usands of different architectures to be trained from scratch. Recent work has ex plored weight sharing across models to amortize the cost of training. Although p revious methods reduced the cost of architecture search by orders of magnitude, they remain complex, requiring hypernetworks or reinforcement learning controlle rs. We aim to understand weight sharing for one-shot architecture search. With c areful experimental analysis, we show that it is possible to efficiently identif y promising architectures from a complex search space without either hypernetwor ks or RL.
****

signSGD: Compressed Optimisation for Non-Convex Problems
Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, Animashree Anandkumar

Training large neural networks requires distributing learning across multiple wo
rkers, where the cost of communicating gradients can be a significant bottleneck
. signSGD alleviates this problem by transmitting just the sign of each minibatc
h stochastic gradient. We prove that it can get the best of both worlds: compres
sed gradients and SGD-level convergence rate. The relative ℓ_1/ℓ_2 geome
try of gradients, noise and curvature informs whether signSGD or SGD is theoreti
cally better suited to a particular problem. On the practical side we find that
the momentum counterpart of signSGD is able to match the accuracy and convergenc
e speed of Adam on deep Imagenet models. We extend our theory to the distributed
 setting, where the parameter server uses majority vote to aggregate gradient si
gns from each worker enabling 1-bit compression of worker-server communication i
n both directions. Using a theorem by Gauss we prove that majority vote can achi
eve the same reduction in variance as full precision distributed SGD. Thus, ther
e is great promise for sign-based optimisation schemes to achieve fast communica
tion and fast convergence. Code to reproduce experiments is to be found at https
://github.com/jxbz/signSGD.
****

Distributed Clustering via LSH Based Data Partitioning
Aditya Bhaskara, Maheshakya Wijewardena
Given the importance of clustering in the analysisof large scale data, distribut
ed algorithms for formulations such as k-means, k-median, etc. have been extensi
vely studied. A successful approach here has been the “reduce and merge” paradig
m, in which each machine reduces its input size to {Õ}(k), and this data reducti
on continues (possibly iteratively) until all the data fits on one machine, at w
hich point the problem is solved locally. This approach has the intrinsic bottle
neck that each machine must solve a problem of size $\geq$ k, and needs to commu
nicate at least Ω(k) points to the other machines. We propose a novel dat
a partitioning idea to overcome this bottleneck, and in effect, have different m
achines focus on “finding different clusters”. Under the assumption that we know
 the optimum value of the objective up to a poly(n) factor (arbitrary polynomial
), we establish worst-case approximation guarantees for our method. We see that
our algorithm results in lower communication as well as a near-optimal number of
 ‘rounds’ of computation (in the popular MapReduce framework).
****

Autoregressive Convolutional Neural Networks for Asynchronous Time Series
Mikolaj Binkowski, Gautier Marti, Philippe Donnat
We propose Significance-Offset Convolutional Neural Network, a deep convolutiona
l network architecture for regression of multivariate asynchronous time series.
The model is inspired by standard autoregressive (AR) models and gating mechanis
ms used in recurrent neural networks. It involves an AR-like weighting system, w
here the final predictor is obtained as a weighted sum of adjusted regressors, w
hile the weights are data-dependent functions learnt through a convolutional net
work. The architecture was designed for applications on asynchronous time series
 and is evaluated on such datasets: a hedge fund proprietary dataset of over 2 m
illion quotes for a credit derivative index, an artificially generated noisy aut
oregressive series and UCI household electricity consumption dataset. The propos
ed architecture achieves promising results as compared to convolutional and recu
rrent neural networks.
****

Adaptive Sampled Softmax with Kernel Based Sampling
Guy Blanc, Steffen Rendle
Softmax is the most commonly used output function for multiclass problems and is
 widely used in areas such as vision, natural language processing, and recommend
ation. A softmax model has linear costs in the number of classes which makes it
too expensive for many real-world problems. A common approach to speed up traini
ng involves sampling only some of the classes at each training step. It is known
 that this method is biased and that the bias increases the more the sampling di
stribution deviates from the output distribution. Nevertheless, almost all recen
t work uses simple sampling distributions that require a large sample size to mi
tigate the bias. In this work, we propose a new class of kernel based sampling m

ethods and develop an efficient sampling algorithm. Kernel based sampling adapts to the model as it is trained, thus resulting in low bias. It can also be easily applied to many models because it relies only on the model’s last hidden layer. We empirically study the trade-off of bias, sampling distribution and sample size and show that kernel based sampling results in low bias with few samples.

****

Optimizing the Latent Space of Generative Networks

Piotr Bojanowski, Armand Joulin, David Lopez-Pas, Arthur Szlam

Generative Adversarial Networks (GANs) have achieved remarkable results in the task of generating realistic natural images. In most successful applications, GAN models share two common aspects: solving a challenging saddle point optimization problem, interpreted as an adversarial game between a generator and a discriminator functions; and parameterizing the generator and the discriminator as deep convolutional neural networks. The goal of this paper is to disentangle the contribution of these two factors to the success of GANs. In particular, we introduce Generative Latent Optimization (GLO), a framework to train deep convolutional generators using simple reconstruction losses. Throughout a variety of experiments, we show that GLO enjoys many of the desirable properties of GANs: synthesizing visually-appealing samples, interpolating meaningfully between samples, and performing linear arithmetic with noise vectors; all of this without the adversarial optimization scheme.

****

NetGAN: Generating Graphs via Random Walks

Aleksandar Bojchevski, Oleksandr Shchur, Daniel Zügner, Stephan Günnemann

We propose NetGAN - the first implicit generative model for graphs able to mimic real-world networks. We pose the problem of graph generation as learning the distribution of biased random walks over the input graph. The proposed model is based on a stochastic neural network that generates discrete output samples and is trained using the Wasserstein GAN objective. NetGAN is able to produce graphs that exhibit well-known network patterns without explicitly specifying them in the model definition. At the same time, our model exhibits strong generalization properties, as highlighted by its competitive link prediction performance, despite not being trained specifically for this task. Being the first approach to combine both of these desirable properties, NetGAN opens exciting avenues for further research.

****

A Progressive Batching L-BFGS Method for Machine Learning

Raghu Bollapragada, Jorge Nocedal, Dheevatsa Mudigere, Hao-Jun Shi, Ping Tak Peter Tang

The standard L-BFGS method relies on gradient approximations that are not dominated by noise, so that search directions are descent directions, the line search is reliable, and quasi-Newton updating yields useful quadratic models of the objective function. All of this appears to call for a full batch approach, but since small batch sizes give rise to faster algorithms with better generalization properties, L-BFGS is currently not considered an algorithm of choice for large-scale machine learning applications. One need not, however, choose between the two extremes represented by the full batch or highly stochastic regimes, and may instead follow a progressive batching approach in which the sample size increases during the course of the optimization. In this paper, we present a new version of the L-BFGS algorithm that combines three basic components - progressive batching, a stochastic line search, and stable quasi-Newton updating - and that performs well on training logistic regression and deep neural networks. We provide supporting convergence theory for the method.

****

Prediction Rule Reshaping

Matt Bonakdarpour, Sabyasachi Chatterjee, Rina Foygel Barber, John Lafferty

Two methods are proposed for high-dimensional shape-constrained regression and classification. These methods reshape pre-trained prediction rules to satisfy shape constraints like monotonicity and convexity. The first method can be applied to any pre-trained prediction rule, while the second method deals specifically w

ith random forests. In both cases, efficient algorithms are developed for comput ing the estimators, and experiments are performed to demonstrate their performan ce on four datasets. We find that reshaping methods enforce shape constraints wi thout compromising predictive accuracy.
****

QuantTree: Histograms for Change Detection in Multivariate Data Streams
Giacomo Boracchi, Diego Carrera, Cristiano Cervellera, Danilo Macciò
We address the problem of detecting distribution changes in multivariate data st reams by means of histograms. Histograms are very general and flexible models, w hich have been relatively ignored in the change-detection literature as they oft en require a number of bins that grows unfeasibly with the data dimension. We pr esent QuantTree, a recursive binary splitting scheme that adaptively defines the histogram bins to ease the detection of any distribution change. Our design sch eme implies that i) we can easily control the overall number of bins and ii) the bin probabilities do not depend on the distribution of stationary data. This la tter is a very relevant aspect in change detection, since thresholds of tests st atistics based on these histograms (e.g., the Pearson statistic or the total var iation) can be numerically computed from univariate and synthetically generated data, yet guaranteeing a controlled false positive rate. Our experiments show th at the proposed histograms are very effective in detecting changes in high dimen sional data streams, and that the resulting thresholds can effectively control t he false positive rate, even when the number of training samples is relatively s mall.
****

Matrix Norms in Data Streams: Faster, Multi-Pass and Row-Order
Vladimir Braverman, Stephen Chestnut, Robert Krauthgamer, Yi Li, David Woodruff, Lin Yang
A central problem in mining massive data streams is characterizing which functio ns of an underlying frequency vector can be approximated efficiently. Given the prevalence of large scale linear algebra problems in machine learning, recently there has been considerable effort in extending this data stream problem to that of estimating functions of a matrix. This setting generalizes classical problem s to the analogous ones for matrices. For example, instead of estimating frequen t-item counts, we now wish to estimate “frequent-direction” counts. A related ex ample is to estimate norms, which now correspond to estimating a vector norm on the singular values of the matrix. Despite recent efforts, the current understan ding for such matrix problems is considerably weaker than that for vector proble ms. We study a number of aspects of estimating matrix norms in a stream that hav e not previously been considered: (1) multi-pass algorithms, (2) algorithms that see the underlying matrix one row at a time, and (3) time-efficient algorithms. Our multi-pass and row-order algorithms use less memory than what is provably r equired in the single-pass and entrywise-update models, and thus give separation s between these models (in terms of memory). Moreover, all of our algorithms are considerably faster than previous ones. We also prove a number of lower bounds, and obtain for instance, a near-complete characterization of the memory require d of row-order algorithms for estimating Schatten p-norms of sparse matrices. We complement our results with numerical experiments.
****

Predict and Constrain: Modeling Cardinality in Deep Structured Prediction
Nataly Brukhim, Amir Globerson
Many machine learning problems require the prediction of multi-dimensional label s. Such structured prediction models can benefit from modeling dependencies betw een labels. Recently, several deep learning approaches to structured prediction have been proposed. Here we focus on capturing cardinality constraints in such m odels. Namely, constraining the number of non-zero labels that the model outputs . Such constraints have proven very useful in previous structured prediction met hods, but it is a challenge to introduce them into a deep learning approach. Her e we show how to do this via a novel deep architecture. Our approach outperforms strong baselines, achieving state-of-the-art results on multi-label classificat ion benchmarks.

****
Quasi-Monte Carlo Variational Inference
Alexander Buchholz, Florian Wenzel, Stephan Mandt
Many machine learning problems involve Monte Carlo gradient estimators. As a pro minent example, we focus on Monte Carlo variational inference (MCVI) in this pap er. The performance of MCVI crucially depends on the variance of its stochastic gradients. We propose variance reduction by means of Quasi-Monte Carlo (QMC) sam pling. QMC replaces N i.i.d. samples from a uniform probability distribution by a deterministic sequence of samples of length N. This sequence covers the underl ying random variable space more evenly than i.i.d. draws, reducing the variance of the gradient estimator. With our novel approach, both the score function and the reparameterization gradient estimators lead to much faster convergence. We a lso propose a new algorithm for Monte Carlo objectives, where we operate with a constant learning rate and increase the number of QMC samples per iteration. We prove that this way, our algorithm can converge asymptotically at a faster rate than SGD . We furthermore provide theoretical guarantees on qmc for Monte Carlo objectives that go beyond MCVI , and support our findings by several experiments on large-scale data sets from various domains.
****
Path-Level Network Transformation for Efficient Architecture Search
Han Cai, Jiacheng Yang, Weinan Zhang, Song Han, Yong Yu
We introduce a new function-preserving transformation for efficient neural archi tecture search. This network transformation allows reusing previously trained ne tworks and existing successful architectures that improves sample efficiency. We aim to address the limitation of current network transformation operations that can only perform layer-level architecture modifications, such as adding (prunin g) filters or inserting (removing) a layer, which fails to change the topology o f connection paths. Our proposed path-level transformation operations enable the meta-controller to modify the path topology of the given network while keeping the merits of reusing weights, and thus allow efficiently designing effective st ructures with complex path topologies like Inception models. We further propose a bidirectional tree-structured reinforcement learning meta-controller to explor e a simple yet highly expressive tree-structured architecture space that can be viewed as a generalization of multi-branch architectures. We experimented on the image classification datasets with limited computational resources (about 200 G PU-hours), where we observed improved parameter efficiency and better test resul ts (97.70% test accuracy on CIFAR-10 with 14.3M parameters and 74.6% top-1 accur acy on ImageNet in the mobile setting), demonstrating the effectiveness and tran sferability of our designed architectures.
****
Improved large-scale graph learning through ridge spectral sparsification
Daniele Calandriello, Alessandro Lazaric, Ioannis Koutis, Michal Valko
The representation and learning benefits of methods based on graph Laplacians, s uch as Laplacian smoothing or harmonic function solution for semi-supervised lea rning (SSL), are empirically and theoretically well supported. Nonetheless, the exact versions of these methods scale poorly with the number of nodes n of the graph. In this paper, we combine a spectral sparsification routine with Laplaci an learning. Given a graph G as input, our algorithm computes a sparsifier in a distributed way in $O(n\log^3(n))$ time, $O(m\log^3(n))$ work and $O(n\log(n)) $ memory, using only $\log(n)$ rounds of communication. Furthermore, motivated b y the regularization often employed in learning algorithms, we show that constru cting sparsifiers that preserve the spectrum of the Laplacian only up to the reg ularization level may drastically reduce the size of the final graph. By constru cting a spectrally-similar graph, we are able to bound the error induced by the sparsification for a variety of downstream tasks (e.g., SSL). We empirically val idate the theoretical guarantees on Amazon co-purchase graph and compare to the state-of-the-art heuristics.
****
Bayesian Coreset Construction via Greedy Iterative Geodesic Ascent
Trevor Campbell, Tamara Broderick

Coherent uncertainty quantification is a key strength of Bayesian methods. But m
odern algorithms for approximate Bayesian posterior inference often sacrifice ac
curate posterior uncertainty estimation in the pursuit of scalability. This work
 shows that previous Bayesian coreset construction algorithms—which build a smal
l, weighted subset of the data that approximates the full dataset—are no excepti
on. We demonstrate that these algorithms scale the coreset log-likelihood subopt
imally, resulting in underestimated posterior uncertainty. To address this short
coming, we develop greedy iterative geodesic ascent (GIGA), a novel algorithm fo
r Bayesian coreset construction that scales the coreset log-likelihood optimally
. GIGA provides geometric decay in posterior approximation error as a function o
f coreset size, and maintains the fast running time of its predecessors. The pap
er concludes with validation of GIGA on both synthetic and real datasets, demons
trating that it reduces posterior approximation error by orders of magnitude com
pared with previous coreset constructions.
****

Adversarial Learning with Local Coordinate Coding
Jiezhang Cao, Yong Guo, Qingyao Wu, Chunhua Shen, Junzhou Huang, Mingkui Tan
Generative adversarial networks (GANs) aim to generate realistic data from some
prior distribution (e.g., Gaussian noises). However, such prior distribution is
often independent of real data and thus may lose semantic information (e.g., geo
metric structure or content in images) of data. In practice, the semantic inform
ation might be represented by some latent distribution learned from data, which,
 however, is hard to be used for sampling in GANs. In this paper, rather than sa
mpling from the pre-defined prior distribution, we propose a Local Coordinate Co
ding (LCC) based sampling method to improve GANs. We derive a generalization bou
nd for LCC based GANs and prove that a small dimensional input is sufficient to
achieve good generalization. Extensive experiments on various real-world dataset
s demonstrate the effectiveness of the proposed method.
****

Fair and Diverse DPP-Based Data Summarization
Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, Nis
heeth Vishnoi
Sampling methods that choose a subset of the data proportional to its diversity
in the feature space are popular for data summarization. However, recent studies
 have noted the occurrence of bias {–} e.g., under or over representation of a p
articular gender or ethnicity {–} in such data summarization methods. In this pa
per we initiate a study of the problem of outputting a diverse and fair summary
of a given dataset. We work with a well-studied determinantal measure of diversi
ty and corresponding distributions (DPPs) and present a framework that allows us
 to incorporate a general class of fairness constraints into such distributions.
 Designing efficient algorithms to sample from these constrained determinantal d
istributions, however, suffers from a complexity barrier; we present a fast samp
ler that is provably good when the input vectors satisfy a natural property. Our
 empirical results on both real-world and synthetic datasets show that the diver
sity of the samples produced by adding fairness constraints is not too far from
the unconstrained case.
****

Conditional Noise-Contrastive Estimation of Unnormalised Models
Ciwan Ceylan, Michael U. Gutmann
Many parametric statistical models are not properly normalised and only specifie
d up to an intractable partition function, which renders parameter estimation di
fficult. Examples of unnormalised models are Gibbs distributions, Markov random
fields, and neural network models in unsupervised deep learning. In previous wor
k, the estimation principle called noise-contrastive estimation (NCE) was introd
uced where unnormalised models are estimated by learning to distinguish between
data and auxiliary noise. An open question is how to best choose the auxiliary n
oise distribution. We here propose a new method that addresses this issue. The p
roposed method shares with NCE the idea of formulating density estimation as a s
upervised learning problem but in contrast to NCE, the proposed method leverages
 the observed data when generating noise samples. The noise can thus be generate

d in a semi-automated manner. We first present the underlying theory of the new method, show that score matching emerges as a limiting case, validate the method on continuous and discrete valued synthetic data, and show that we can expect an improved performance compared to NCE when the data lie in a lower-dimensional manifold. Then we demonstrate its applicability in unsupervised deep learning by estimating a four-layer neural image model.

****

Adversarial Time-to-Event Modeling

Paidamoyo Chapfuwa, Chenyang Tao, Chunyuan Li, Courtney Page, Benjamin Goldstein, Lawrence Carin Duke, Ricardo Henao

Modern health data science applications leverage abundant molecular and electronic health data, providing opportunities for machine learning to build statistical models to support clinical practice. Time-to-event analysis, also called survival analysis, stands as one of the most representative examples of such statistical models. We present a deep-network-based approach that leverages adversarial learning to address a key challenge in modern time-to-event modeling: nonparametric estimation of event-time distributions. We also introduce a principled cost function to exploit information from censored events (events that occur subsequent to the observation window). Unlike most time-to-event models, we focus on the estimation of time-to-event distributions, rather than time ordering. We validate our model on both benchmark and real datasets, demonstrating that the proposed formulation yields significant performance gains relative to a parametric alternative, which we also propose.

****

Stability and Generalization of Learning Algorithms that Converge to Global Optima

Zachary Charles, Dimitris Papailiopoulos

We establish novel generalization bounds for learning algorithms that converge to global minima. We derive black-box stability results that only depend on the convergence of a learning algorithm and the geometry around the minimizers of the empirical risk function. The results are shown for non-convex loss functions satisfying the Polyak-Lojasiewicz (PL) and the quadratic growth (QG) conditions, which we show arise for 1-layer neural networks with leaky ReLU activations and deep neural networks with linear activations. We use our results to establish the stability of first-order methods such as stochastic gradient descent (SGD), gradient descent (GD), randomized coordinate descent (RCD), and the stochastic variance reduced gradient method (SVRG), in both the PL and the strongly convex setting. Our results match or improve state-of-the-art generalization bounds and can easily extend to similar optimization algorithms. Finally, although our results imply comparable stability for SGD and GD in the PL setting, we show that there exist simple quadratic models with multiple local minima where SGD is stable but GD is not.

****

Learning and Memorization

Satrajit Chatterjee

In the machine learning research community, it is generally believed that there is a tension between memorization and generalization. In this work we examine to what extent this tension exists by exploring if it is possible to generalize by memorizing alone. Although direct memorization with a lookup table obviously does not generalize, we find that introducing depth in the form of a network of support-limited lookup tables leads to generalization that is significantly above chance and closer to those obtained by standard learning algorithms on several tasks derived from MNIST and CIFAR-10. Furthermore, we demonstrate through a series of empirical results that our approach allows for a smooth tradeoff between memorization and generalization and exhibits some of the most salient characteristics of neural networks: depth improves performance; random data can be memorized and yet there is generalization on real data; and memorizing random data is harder in a certain sense than memorizing real data. The extreme simplicity of the algorithm and potential connections with generalization theory point to several interesting directions for future research.

****
On the Theory of Variance Reduction for Stochastic Gradient Monte Carlo
Niladri Chatterji, Nicolas Flammarion, Yian Ma, Peter Bartlett, Michael Jordan
We provide convergence guarantees in Wasserstein distance for a variety of variance-reduction methods: SAGA Langevin diffusion, SVRG Langevin diffusion and control-variate underdamped Langevin diffusion. We analyze these methods under a uniform set of assumptions on the log-posterior distribution, assuming it to be smooth, strongly convex and Hessian Lipschitz. This is achieved by a new proof technique combining ideas from finite-sum optimization and the analysis of sampling methods. Our sharp theoretical bounds allow us to identify regimes of interest where each method performs better than the others. Our theory is verified with experiments on real-world and synthetic datasets.
****
Hierarchical Clustering with Structural Constraints
Vaggos Chatziafratis, Rad Niazadeh, Moses Charikar
Hierarchical clustering is a popular unsupervised data analysis method. For many real-world applications, we would like to exploit prior information about the data that imposes constraints on the clustering hierarchy, and is not captured by the set of features available to the algorithm. This gives rise to the problem of hierarchical clustering with structural constraints. Structural constraints pose major challenges for bottom-up approaches like average/single linkage and even though they can be naturally incorporated into top-down divisive algorithms, no formal guarantees exist on the quality of their output. In this paper, we provide provable approximation guarantees for two simple top-down algorithms, using a recently introduced optimization viewpoint of hierarchical clustering with pairwise similarity information (Dasgupta, 2016). We show how to find good solutions even in the presence of conflicting prior information, by formulating a constraint-based regularization of the objective. Furthemore, we explore a variation of this objective for dissimilarity information (Cohen-Addad et al., 2018) and improve upon current techniques. Finally, we demonstrate our approach on a real dataset for the taxonomy application.
****
Hierarchical Deep Generative Models for Multi-Rate Multivariate Time Series
Zhengping Che, Sanjay Purushotham, Guangyu Li, Bo Jiang, Yan Liu
Multi-Rate Multivariate Time Series (MR-MTS) are the multivariate time series observations which come with various sampling rates and encode multiple temporal dependencies. State-space models such as Kalman filters and deep learning models such as deep Markov models are mainly designed for time series data with the same sampling rate and cannot capture all the dependencies present in the MR-MTS data. To address this challenge, we propose the Multi-Rate Hierarchical Deep Markov Model (MR-HDMM), a novel deep generative model which uses the latent hierarchical structure with a learnable switch mechanism to capture the temporal dependencies of MR-MTS. Experimental results on two real-world datasets demonstrate that our MR-HDMM model outperforms the existing state-of-the-art deep learning and state-space models on forecasting and interpolation tasks. In addition, the latent hierarchies in our model provide a way to show and interpret the multiple temporal dependencies.
****
GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks
Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, Andrew Rabinovich
Deep multitask networks, in which one neural network produces multiple predictive outputs, can offer better speed and performance than their single-task counterparts but are challenging to train properly. We present a gradient normalization (GradNorm) algorithm that automatically balances training in deep multitask models by dynamically tuning gradient magnitudes. We show that for various network architectures, for both regression and classification tasks, and on both synthetic and real datasets, GradNorm improves accuracy and reduces overfitting across multiple tasks when compared to single-task networks, static baselines, and other adaptive multitask loss balancing techniques. GradNorm also matches or surpass

es the performance of exhaustive grid search methods, despite only involving a s ingle asymmetry hyperparameter α. Thus, what was once a tedious search pr ocess that incurred exponentially more compute for each task added can now be ac complished within a few training runs, irrespective of the number of tasks. Ulti mately, we will demonstrate that gradient manipulation affords us great control over the training dynamics of multitask networks and may be one of the keys to u nlocking the potential of multitask learning.
****
Weakly Submodular Maximization Beyond Cardinality Constraints: Does Randomizatio n Help Greedy?
Lin Chen, Moran Feldman, Amin Karbasi
Submodular functions are a broad class of set functions that naturally arise in many machine learning applications. Due to their combinatorial structures, there has been a myriad of algorithms for maximizing such functions under various con straints. Unfortunately, once a function deviates from submodularity (even sligh tly), the known algorithms may perform arbitrarily poorly. Amending this issue, by obtaining approximation results for functions obeying properties that general ize submodularity, has been the focus of several recent works. One such class, k nown as weakly submodular functions, has received a lot of recent attention from the machine learning community due to its strong connections to restricted stro ng convexity and sparse reconstruction. In this paper, we prove that a randomize d version of the greedy algorithm achieves an approximation ratio of $(1 + 1/\ga mma)^{-2}$ for weakly submodular maximization subject to a general matroid cons traint, where γ is a parameter measuring the distance from submodularity. To the best of our knowledge, this is the first algorithm with a non-trivial ap proximation guarantee for this constrained optimization problem. Moreover, our e xperimental results show that our proposed algorithm performs well in a variety of real-world problems, including regression, video summarization, splice site d etection, and black-box interpretation.
****
Projection-Free Online Optimization with Stochastic Gradient: From Convexity to Submodularity
Lin Chen, Christopher Harshaw, Hamed Hassani, Amin Karbasi
Online optimization has been a successful framework for solving large-scale prob lems under computational constraints and partial information. Current methods fo r online convex optimization require either a projection or exact gradient compu tation at each step, both of which can be prohibitively expensive for large-scal e applications. At the same time, there is a growing trend of non-convex optimiz ation in machine learning community and a need for online methods. Continuous DR -submodular functions, which exhibit a natural diminishing returns condition, ha ve recently been proposed as a broad class of non-convex functions which may be efficiently optimized. Although online methods have been introduced, they suffer from similar problems. In this work, we propose Meta-Frank-Wolfe, the first onl ine projection-free algorithm that uses stochastic gradient estimates. The algor ithm relies on a careful sampling of gradients in each round and achieves the op timal $O(\sqrt{T})$ adversarial regret bounds for convex and continuous submodu lar optimization. We also propose One-Shot Frank-Wolfe, a simpler algorithm whic h requires only a single stochastic gradient estimate in each round and achieves an $O(T^{2/3})$ stochastic regret bound for convex and continuous submodular op timization. We apply our methods to develop a novel "lifting" framework for the online discrete submodular maximization and also see that they outperform curren t state-of-the-art techniques on various experiments.
****
Continuous-Time Flows for Efficient Inference and Density Estimation
Changyou Chen, Chunyuan Li, Liqun Chen, Wenlin Wang, Yunchen Pu, Lawrence Carin Duke
Two fundamental problems in unsupervised learning are efficient inference for la tent-variable models and robust density estimation based on large amounts of unl abeled data. Algorithms for the two tasks, such as normalizing flows and generat ive adversarial networks (GANs), are often developed independently. In this pape

r, we propose the concept of continuous-time flows (CTFs), a family of diffusion -based methods that are able to asymptotically approach a target distribution. D istinct from normalizing flows and GANs, CTFs can be adopted to achieve the abov e two goals in one framework, with theoretical guarantees. Our framework include s distilling knowledge from a CTF for efficient inference, and learning an expli cit energy-based distribution with CTFs for density estimation. Both tasks rely on a new technique for distribution matching within amortized learning. Experime nts on various tasks demonstrate promising performance of the proposed CTF frame work, compared to related techniques.

****

Scalable Bilinear Pi Learning Using State and Action Features

Yichen Chen, Lihong Li, Mengdi Wang

Approximate linear programming (ALP) represents one of the major algorithmic fam ilies to solve large-scale Markov decision processes (MDP). In this work, we stu dy a primal-dual formulation of the ALP, and develop a scalable, model-free algo rithm called bilinear π learning for reinforcement learning when a sampling oracle is provided. This algorithm enjoys a number of advantages. First, it adop ts linear and bilinear models to represent the high-dimensional value function a nd state-action distributions, respectively, using given state and action featur es. Its run-time complexity depends on the number of features, not the size of t he underlying MDPs. Second, it operates in a fully online fashion without having to store any sample, thus having minimal memory footprint. Third, we prove that it is sample-efficient, solving for the optimal policy to high precision with a sample complexity linear in the dimension of the parameter space.

****

Stein Points

Wilson Ye Chen, Lester Mackey, Jackson Gorham, Francois-Xavier Briol, Chris Oate s

An important task in computational statistics and machine learning is to approxi mate a posterior distribution $p(x)$ with an empirical measure supported on a se t of representative points $\{x_i\}_{i=1}^n$. This paper focuses on methods wher e the selection of points is essentially deterministic, with an emphasis on achi eving accurate approximation when n is small. To this end, we present Stein Po ints. The idea is to exploit either a greedy or a conditional gradient method to iteratively minimise a kernel Stein discrepancy between the empirical measure a nd $p(x)$. Our empirical results demonstrate that Stein Points enable accurate a pproximation of the posterior at modest computational cost. In addition, theoret ical results are provided to establish convergence of the method.

****

Learning K-way D-dimensional Discrete Codes for Compact Embedding Representation s

Ting Chen, Martin Renqiang Min, Yizhou Sun

Conventional embedding methods directly associate each symbol with a continuous embedding vector, which is equivalent to applying a linear transformation based on a “one-hot” encoding of the discrete symbols. Despite its simplicity, such ap proach yields the number of parameters that grows linearly with the vocabulary s ize and can lead to overfitting. In this work, we propose a much more compact K- way D-dimensional discrete encoding scheme to replace the “one-hot" encoding. In the proposed “KD encoding”, each symbol is represented by a D-dimensional cod e with a cardinality of K, and the final symbol embedding vector is generated by composing the code embedding vectors. To end-to-end learn semantically meanin gful codes, we derive a relaxed discrete optimization approach based on stochast ic gradient descent, which can be generally applied to any differentiable comput ational graph with an embedding layer. In our experiments with various applicati ons from natural language processing to graph convolutional networks, the total size of the embedding layer can be reduced up to 98% while achieving similar or better performance.

****

PixelSNAIL: An Improved Autoregressive Generative Model

XI Chen, Nikhil Mishra, Mostafa Rohaninejad, Pieter Abbeel

Autoregressive generative models achieve the best results in density estimation tasks involving high dimensional data, such as images or audio. They pose density estimation as a sequence modeling task, where a recurrent neural network (RNN) models the conditional distribution over the next element conditioned on all previous elements. In this paradigm, the bottleneck is the extent to which the RNN can model long-range dependencies, and the most successful approaches rely on causal convolutions. Taking inspiration from recent work in meta reinforcement learning, where dealing with long-range dependencies is also essential, we introduce a new generative model architecture that combines causal convolutions with self attention. In this paper, we describe the resulting model and present state-of-the-art log-likelihood results on heavily benchmarked datasets: CIFAR-10, 32×32 ImageNet and 64×64 ImageNet. Our implementation will be made available at \url{https://github.com/neocxi/pixelsnail-public}.

****

Dynamical Isometry and a Mean Field Theory of RNNs: Gating Enables Signal Propagation in Recurrent Neural Networks

Minmin Chen, Jeffrey Pennington, Samuel Schoenholz

Recurrent neural networks have gained widespread use in modeling sequence data across various domains. While many successful recurrent architectures employ a notion of gating, the exact mechanism that enables such remarkable performance is not well understood. We develop a theory for signal propagation in recurrent networks after random initialization using a combination of mean field theory and random matrix theory. To simplify our discussion, we introduce a new RNN cell with a simple gating mechanism that we call the minimalRNN and compare it with vanilla RNNs. Our theory allows us to define a maximum timescale over which RNNs can remember an input. We show that this theory predicts trainability for both recurrent architectures. We show that gated recurrent networks feature a much broader, more robust, trainable region than vanilla RNNs, which corroborates recent experimental findings. Finally, we develop a closed-form critical initialization scheme that achieves dynamical isometry in both vanilla RNNs and minimalRNNs. We show that this results in significantly improved training dynamics. Finally, we demonstrate that the minimalRNN achieves comparable performance to its more complex counterparts, such as LSTMs or GRUs, on a language modeling task.

****

Learning to Explain: An Information-Theoretic Perspective on Model Interpretation

Jianbo Chen, Le Song, Martin Wainwright, Michael Jordan

We introduce instancewise feature selection as a methodology for model interpretation. Our method is based on learning a function to extract a subset of features that are most informative for each given example. This feature selector is trained to maximize the mutual information between selected features and the response variable, where the conditional distribution of the response variable given the input is the model to be explained. We develop an efficient variational approximation to the mutual information, and show the effectiveness of our method on a variety of synthetic and real data sets using both quantitative metrics and human evaluation.

****

Variational Inference and Model Selection with Generalized Evidence Bounds

Liqun Chen, Chenyang Tao, Ruiyi Zhang, Ricardo Henao, Lawrence Carin Duke

Recent advances on the scalability and flexibility of variational inference have made it successful at unravelling hidden patterns in complex data. In this work we propose a new variational bound formulation, yielding an estimator that extends beyond the conventional variational bound. It naturally subsumes the importance-weighted and Renyi bounds as special cases, and it is provably sharper than these counterparts. We also present an improved estimator for variational learning, and advocate a novel high signal-to-variance ratio update rule for the variational parameters. We discuss model-selection issues associated with existing evidence-lower-bound-based variational inference procedures, and show how to leverage the flexibility of our new formulation to address them. Empirical evidence is provided to validate our claims.

****
DRACO: Byzantine-resilient Distributed Training via Redundant Gradients
Lingjiao Chen, Hongyi Wang, Zachary Charles, Dimitris Papailiopoulos
Distributed model training is vulnerable to byzantine system failures and adversarial compute nodes, i.e., nodes that use malicious updates to corrupt the global model stored at a parameter server (PS). To guarantee some form of robustness, recent work suggests using variants of the geometric median as an aggregation rule, in place of gradient averaging. Unfortunately, median-based rules can incur a prohibitive computational overhead in large-scale settings, and their convergence guarantees often require strong assumptions. In this work, we present DRACO, a scalable framework for robust distributed training that uses ideas from coding theory. In DRACO, each compute node evaluates redundant gradients that are used by the parameter server to eliminate the effects of adversarial updates. DRACO comes with problem-independent robustness guarantees, and the model that it trains is identical to the one trained in the adversary-free setup. We provide extensive experiments on real datasets and distributed setups across a variety of large-scale models, where we show that DRACO is several times, to orders of magnitude faster than median-based approaches.
****
SADAGRAD: Strongly Adaptive Stochastic Gradient Methods
Zaiyi Chen, Yi Xu, Enhong Chen, Tianbao Yang
Although the convergence rates of existing variants of ADAGRAD have a better dependence on the number of iterations under the strong convexity condition, their iteration complexities have a explicitly linear dependence on the dimensionality of the problem. To alleviate this bad dependence, we propose a simple yet novel variant of ADAGRAD for stochastic (weakly) strongly convex optimization. Different from existing variants, the proposed variant (referred to as SADAGRAD) uses an adaptive restarting scheme in which (i) ADAGRAD serves as a sub-routine and is restarted periodically; (ii) the number of iterations for restarting ADAGRAD depends on the history of learning that incorporates knowledge of the geometry of the data. In addition to the adaptive proximal functions and adaptive number of iterations for restarting, we also develop a variant that is adaptive to the (implicit) strong convexity from the data, which together makes the proposed algorithm strongly adaptive. In terms of iteration complexity, in the worst case SADAGRAD has an O(1/\epsilon) for finding an \epsilon-optimal solution similar to other variants. However, it could enjoy faster convergence and much better dependence on the problem’s dimensionality when stochastic gradients are sparse. Extensive experiments on large-scale data sets demonstrate the efficiency of the proposed algorithms in comparison with several variants of ADAGRAD and stochastic gradient method.
****
Covariate Adjusted Precision Matrix Estimation via Nonconvex Optimization
Jinghui Chen, Pan Xu, Lingxiao Wang, Jian Ma, Quanquan Gu
We propose a nonconvex estimator for the covariate adjusted precision matrix estimation problem in the high dimensional regime, under sparsity constraints. To solve this estimator, we propose an alternating gradient descent algorithm with hard thresholding. Compared with existing methods along this line of research, which lack theoretical guarantees in optimization error and/or statistical error, the proposed algorithm not only is computationally much more efficient with a linear rate of convergence, but also attains the optimal statistical rate up to a logarithmic factor. Thorough experiments on both synthetic and real data support our theory.
****
End-to-End Learning for the Deep Multivariate Probit Model
Di Chen, Yexiang Xue, Carla Gomes
The multivariate probit model (MVP) is a popular classic model for studying binary responses of multiple entities. Nevertheless, the computational challenge of learning the MVP model, given that its likelihood involves integrating over a multidimensional constrained space of latent variables, significantly limits its application in practice. We propose a flexible deep generalization of the classic

MVP, the Deep Multivariate Probit Model (DMVP), which is an end-to-end learning scheme that uses an efficient parallel sampling process of the multivariate probit model to exploit GPU-boosted deep neural networks. We present both theoretical and empirical analysis of the convergence behavior of DMVP's sampling process with respect to the resolution of the correlation structure. We provide convergence guarantees for DMVP and our empirical analysis demonstrates the advantages of DMVP's sampling compared with standard MCMC-based methods. We also show that when applied to multi-entity modelling problems, which are natural DMVP applications, DMVP trains faster than classical MVP, by at least an order of magnitude, captures rich correlations among entities, and further improves the joint likelihood of entities compared with several competitive models.

****

Stochastic Training of Graph Convolutional Networks with Variance Reduction

Jianfei Chen, Jun Zhu, Le Song

Graph convolutional networks (GCNs) are powerful deep neural networks for graph-structured data. However, GCN computes the representation of a node recursively from its neighbors, making the receptive field size grow exponentially with the number of layers. Previous attempts on reducing the receptive field size by subsampling neighbors do not have convergence guarantee, and their receptive field size per node is still in the order of hundreds. In this paper, we develop control variate based algorithms with new theoretical guarantee to converge to a local optimum of GCN regardless of the neighbor sampling size. Empirical results show that our algorithms enjoy similar convergence rate and model quality with the exact algorithm using only two neighbors per node. The running time of our algorithms on a large Reddit dataset is only one seventh of previous neighbor sampling algorithms.

****

Extreme Learning to Rank via Low Rank Assumption

Minhao Cheng, Ian Davidson, Cho-Jui Hsieh

We consider the setting where we wish to perform ranking for hundreds of thousands of users which is common in recommender systems and web search ranking. Learning a single ranking function is unlikely to capture the variability across all users while learning a ranking function for each person is time-consuming and requires large amounts of data from each user. To address this situation, we propose a Factorization RankSVM algorithm which learns a series of k basic ranking functions and then constructs for each user a local ranking function that is a combination of them. We develop a fast algorithm to reduce the time complexity of gradient descent solver by exploiting the low-rank structure, and the resulting algorithm is much faster than existing methods. Furthermore, we prove that the generalization error of the proposed method can be significantly better than training individual RankSVMs. Finally, we present some interesting patterns in the principal ranking functions learned by our algorithms.

****

Learning a Mixture of Two Multinomial Logits

Flavio Chierichetti, Ravi Kumar, Andrew Tomkins

The classical Multinomial Logit (MNL) is a behavioral model for user choice. In this model, a user is offered a slate of choices (a subset of a finite universe of n items), and selects exactly one item from the slate, each with probability proportional to its (positive) weight. Given a set of observed slates and choices, the likelihood-maximizing item weights are easy to learn at scale, and easy to interpret. However, the model fails to represent common real-world behavior. As a result, researchers in user choice often turn to mixtures of MNLs, which are known to approximate a large class of models of rational user behavior. Unfortunately, the only known algorithms for this problem have been heuristic in nature. In this paper we give the first polynomial-time algorithms for exact learning of uniform mixtures of two MNLs. Interestingly, the parameters of the model can be learned for any n by sampling the behavior of random users only on slates of sizes 2 and 3; in contrast, we show that slates of size 2 are insufficient by themselves.

****

Structured Evolution with Compact Architectures for Scalable Policy Optimization
Krzysztof Choromanski, Mark Rowland, Vikas Sindhwani, Richard Turner, Adrian Weller

We present a new method of blackbox optimization via gradient approximation with the use of structured random orthogonal matrices, providing more accurate estimators than baselines and with provable theoretical guarantees. We show that this algorithm can be successfully applied to learn better quality compact policies than those using standard gradient estimation techniques. The compact policies we learn have several advantages over unstructured ones, including faster training algorithms and faster inference. These benefits are important when the policy is deployed on real hardware with limited resources. Further, compact policies provide more scalable architectures for derivative-free optimization (DFO) in high-dimensional spaces. We show that most robotics tasks from the OpenAI Gym can be solved using neural networks with less than 300 parameters, with almost linear time complexity of the inference phase, with up to 13x fewer parameters relative to the Evolution Strategies (ES) algorithm introduced by Salimans et al. (2017). We do not need heuristics such as fitness shaping to learn good quality policies, resulting in a simple and theoretically motivated training mechanism.

****

Path Consistency Learning in Tsallis Entropy Regularized MDPs
Yinlam Chow, Ofir Nachum, Mohammad Ghavamzadeh

We study the sparse entropy-regularized reinforcement learning (ERL) problem in which the entropy term is a special form of the Tsallis entropy. The optimal policy of this formulation is sparse, i.e., at each state, it has non-zero probability for only a small number of actions. This addresses the main drawback of the standard Shannon entropy-regularized RL (soft ERL) formulation, in which the optimal policy is softmax, and thus, may assign a non-negligible probability mass to non-optimal actions. This problem is aggravated as the number of actions is increased. In this paper, we follow the work of Nachum et al. (2017) in the soft ERL setting, and propose a class of novel path consistency learning (PCL) algorithms, called sparse PCL, for the sparse ERL problem that can work with both on-policy and off-policy data. We first derive a sparse consistency equation that specifies a relationship between the optimal value function and policy of the sparse ERL along any system trajectory. Crucially, a weak form of the converse is also true, and we quantify the sub-optimality of a policy which satisfies sparse consistency, and show that as we increase the number of actions, this sub-optimality is better than that of the soft ERL optimal policy. We then use this result to derive the sparse PCL algorithms. We empirically compare sparse PCL with its soft counterpart, and show its advantage, especially in problems with a large number of actions.

****

An Iterative, Sketching-based Framework for Ridge Regression
Agniva Chowdhury, Jiasen Yang, Petros Drineas

Ridge regression is a variant of regularized least squares regression that is particularly suitable in settings where the number of predictor variables greatly exceeds the number of observations. We present a simple, iterative, sketching-based algorithm for ridge regression that guarantees high-quality approximations to the optimal solution vector. Our analysis builds upon two simple structural results that boil down to randomized matrix multiplication, a fundamental and well-understood primitive of randomized linear algebra. An important contribution of our work is the analysis of the behavior of subsampled ridge regression problems when the ridge leverage scores are used: we prove that accurate approximations can be achieved by a sample whose size depends on the degrees of freedom of the ridge-regression problem rather than the dimensions of the design matrix. Our experimental evaluations verify our theoretical results on both real and synthetic data.

****

Stochastic Wasserstein Barycenters
Sebastian Claici, Edward Chien, Justin Solomon

We present a stochastic algorithm to compute the barycenter of a set of probabil

ity distributions under the Wasserstein metric from optimal transport. Unlike previous approaches, our method extends to continuous input distributions and allows the support of the barycenter to be adjusted in each iteration. We tackle the problem without regularization, allowing us to recover a sharp output whose support is contained within the support of the true barycenter. We give examples where our algorithm recovers a more meaningful barycenter than previous work. Our method is versatile and can be extended to applications such as generating super samples from a given distribution and recovering blue noise approximations.
****

Self-Consistent Trajectory Autoencoder: Hierarchical Reinforcement Learning with Trajectory Embeddings
John Co-Reyes, YuXuan Liu, Abhishek Gupta, Benjamin Eysenbach, Pieter Abbeel, Sergey Levine
In this work, we take a representation learning perspective on hierarchical reinforcement learning, where the problem of learning lower layers in a hierarchy is transformed into the problem of learning trajectory-level generative models. We show that we can learn continuous latent representations of trajectories, which are effective in solving temporally extended and multi-stage problems. Our proposed model, SeCTAR, draws inspiration from variational autoencoders, and learns latent representations of trajectories. A key component of this method is to learn both a latent-conditioned policy and a latent-conditioned model which are consistent with each other. Given the same latent, the policy generates a trajectory which should match the trajectory predicted by the model. This model provides a built-in prediction mechanism, by predicting the outcome of closed loop policy behavior. We propose a novel algorithm for performing hierarchical RL with this model, combining model-based planning in the learned latent space with an unsupervised exploration objective. We show that our model is effective at reasoning over long horizons with sparse rewards for several simulated tasks, outperforming standard reinforcement learning methods and prior methods for hierarchical reasoning, model-based planning, and exploration. This model provides a built-in prediction mechanism, by predicting the outcome of closed loop policy behavior. We propose a novel algorithm for performing hierarchical RL with this model, combining model-based planning in the learned latent space with an unsupervised exploration objective. We show that our model is effective at reasoning over long horizons with sparse rewards for several simulated tasks, outperforming standard reinforcement learning methods and prior methods for hierarchical reasoning, model-based planning, and exploration.
****

On Acceleration with Noise-Corrupted Gradients
Michael Cohen, Jelena Diakonikolas, Lorenzo Orecchia
Accelerated algorithms have broad applications in large-scale optimization, due to their generality and fast convergence. However, their stability in the practical setting of noise-corrupted gradient oracles is not well-understood. This paper provides two main technical contributions: (i) a new accelerated method AGDP that generalizes Nesterov’s AGD and improves on the recent method AXGD (Diakonikolas & Orecchia, 2018), and (ii) a theoretical study of accelerated algorithms under noisy and inexact gradient oracles, which is supported by numerical experiments. This study leverages the simplicity of AGDP and its analysis to clarify the interaction between noise and acceleration and to suggest modifications to the algorithm that reduce the mean and variance of the error incurred due to the gradient noise.
****

Online Linear Quadratic Control
Alon Cohen, Avinatan Hasidim, Tomer Koren, Nevena Lazic, Yishay Mansour, Kunal Talwar
We study the problem of controlling linear time-invariant systems with known noisy dynamics and adversarially chosen quadratic losses. We present the first efficient online learning algorithms in this setting that guarantee $O(\sqrt{T})$ regret under mild assumptions, where T is the time horizon. Our algorithms rely on a novel SDP relaxation for the steady-state distribution of the system. Cruci

ally, and in contrast to previously proposed relaxations, the feasible solutions of our SDP all correspond to “strongly stable” policies that mix exponentially fast to a steady state.
****

GEP-PG: Decoupling Exploration and Exploitation in Deep Reinforcement Learning Algorithms

Cédric Colas, Olivier Sigaud, Pierre-Yves Oudeyer

In continuous action domains, standard deep reinforcement learning algorithms like DDPG suffer from inefficient exploration when facing sparse or deceptive reward problems. Conversely, evolutionary and developmental methods focusing on exploration like Novelty Search, Quality-Diversity or Goal Exploration Processes explore more robustly but are less efficient at fine-tuning policies using gradient-descent. In this paper, we present the GEP-PG approach, taking the best of both worlds by sequentially combining a Goal Exploration Process and two variants of DDPG . We study the learning performance of these components and their combination on a low dimensional deceptive reward problem and on the larger Half-Cheetah benchmark. We show that DDPG fails on the former and that GEP-PG improves over the best DDPG variant in both environments.
****

Efficient Model-Based Deep Reinforcement Learning with Variational State Tabulation

Dane Corneil, Wulfram Gerstner, Johanni Brea

Modern reinforcement learning algorithms reach super-human performance on many board and video games, but they are sample inefficient, i.e. they typically require significantly more playing experience than humans to reach an equal performance level. To improve sample efficiency, an agent may build a model of the environment and use planning methods to update its policy. In this article we introduce Variational State Tabulation (VaST), which maps an environment with a high-dimensional state space (e.g. the space of visual inputs) to an abstract tabular model. Prioritized sweeping with small backups, a highly efficient planning method , can then be used to update state-action values. We show how VaST can rapidly learn to maximize reward in tasks like 3D navigation and efficiently adapt to sudden changes in rewards or transition probabilities.
****

Online Learning with Abstention

Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, Scott Yang

We present an extensive study of a key problem in online learning where the learner can opt to abstain from making a prediction, at a certain cost. In the adversarial setting, we show how existing online algorithms and guarantees can be adapted to this problem. In the stochastic setting, we first point out a bias problem that limits the straightforward extension of algorithms such as UCB-N to this context. Next, we give a new algorithm, UCB-GT, that exploits historical data and time-varying feedback graphs. We show that this algorithm benefits from more favorable regret guarantees than a natural extension of UCB-N . We further report the results of a series of experiments demonstrating that UCB-GT largely outperforms that extension of UCB-N, as well as other standard baselines.
****

Constrained Interacting Submodular Groupings

Andrew Cotter, Mahdi Milani Fard, Seungil You, Maya Gupta, Jeff Bilmes

We introduce the problem of grouping a finite ground set into blocks where each block is a subset of the ground set and where: (i) the blocks are individually highly valued by a submodular function (both robustly and in the average case) while satisfying block-specific matroid constraints; and (ii) block scores interact where blocks are jointly scored highly, thus making the blocks mutually non-redundant. Submodular functions are good models of information and diversity; thus , the above can be seen as grouping the ground set into matroid constrained blocks that are both intra- and inter-diverse. Potential applications include forming ensembles of classification/regression models, partitioning data for parallel processing, and summarization. In the non-robust case, we reduce the problem to non-monotone submodular maximization subject to multiple matroid constraints. In

the mixed robust/average case, we offer a bi-criterion guarantee for a polynomial time deterministic algorithm and a probabilistic guarantee for randomized algorithm, as long as the involved submodular functions (including the inter-block interaction terms) are monotone. We close with a case study in which we use these algorithms to find high quality diverse ensembles of classifiers, showing good results.

****

Inference Suboptimality in Variational Autoencoders

Chris Cremer, Xuechen Li, David Duvenaud

Amortized inference allows latent-variable models trained via variational learning to scale to large datasets. The quality of approximate inference is determined by two factors: a) the capacity of the variational distribution to match the true posterior and b) the ability of the recognition network to produce good variational parameters for each datapoint. We examine approximate inference in variational autoencoders in terms of these factors. We find that divergence from the true posterior is often due to imperfect recognition networks, rather than the limited complexity of the approximating distribution. We show that this is due partly to the generator learning to accommodate the choice of approximation. Furthermore, we show that the parameters used to increase the expressiveness of the approximation play a role in generalizing inference rather than simply improving the complexity of the approximation.

****

Mix & Match Agent Curricula for Reinforcement Learning

Wojciech Czarnecki, Siddhant Jayakumar, Max Jaderberg, Leonard Hasenclever, Yee Whye Teh, Nicolas Heess, Simon Osindero, Razvan Pascanu

We introduce Mix and match (M&M) – a training framework designed to facilitate rapid and effective learning in RL agents that would be too slow or too challenging to train otherwise.The key innovation is a procedure that allows us to automatically form a curriculum over agents. Through such a curriculum we can progressively train more complex agents by, effectively, bootstrapping from solutions found by simpler agents.In contradistinction to typical curriculum learning approaches, we do not gradually modify the tasks or environments presented, but instead use a process to gradually alter how the policy is represented internally.We show the broad applicability of our method by demonstrating significant performance gains in three different experimental setups: (1) We train an agent able to control more than 700 actions in a challenging 3D first-person task; using our method to progress through an action-space curriculum we achieve both faster training and better final performance than one obtains using traditional methods.(2) We further show that M&M can be used successfully to progress through a curriculum of architectural variants defining an agents internal state. (3) Finally, we illustrate how a variant of our method can be used to improve agent performance in a multitask setting.

****

Implicit Quantile Networks for Distributional Reinforcement Learning

Will Dabney, Georg Ostrovski, David Silver, Remi Munos

In this work, we build on recent advances in distributional reinforcement learning to give a generally applicable, flexible, and state-of-the-art distributional variant of DQN. We achieve this by using quantile regression to approximate the full quantile function for the state-action return distribution. By reparameterizing a distribution over the sample space, this yields an implicitly defined return distribution and gives rise to a large class of risk-sensitive policies. We demonstrate improved performance on the 57 Atari 2600 games in the ALE, and use our algorithm’s implicitly defined distributions to study the effects of risk-sensitive policies in Atari games.

****

Learning Steady-States of Iterative Algorithms over Graphs

Hanjun Dai, Zornitsa Kozareva, Bo Dai, Alex Smola, Le Song

Many graph analytics problems can be solved via iterative algorithms where the solutions are often characterized by a set of steady-state conditions. Different algorithms respect to different set of fixed point constraints, so instead of us

ing these traditional algorithms, can we learn an algorithm which can obtain the same steady-state solutions automatically from examples, in an effective and scalable way? How to represent the meta learner for such algorithm and how to carry out the learning? In this paper, we propose an embedding representation for iterative algorithms over graphs, and design a learning method which alternates between updating the embeddings and projecting them onto the steady-state constraints. We demonstrate the effectiveness of our framework using a few commonly used graph algorithms, and show that in some cases, the learned algorithm can handle graphs with more than 100,000,000 nodes in a single machine.

****

Adversarial Attack on Graph Structured Data

Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, Le Song

Deep learning on graph structures has shown exciting results in various applications. However, few attentions have been paid to the robustness of such models, in contrast to numerous research work for image or text adversarial attack and defense. In this paper, we focus on the adversarial attacks that fool deep learning models by modifying the combinatorial structure of data. We first propose a reinforcement learning based attack method that learns the generalizable attack policy, while only requiring prediction labels from the target classifier. We further propose attack methods based on genetic algorithms and gradient descent in the scenario where additional prediction confidence or gradients are available. We use both synthetic and real-world data to show that, a family of Graph Neural Network models are vulnerable to these attacks, in both graph-level and node-level classification tasks. We also show such attacks can be used to diagnose the learned classifiers.

****

SBEED: Convergent Reinforcement Learning with Nonlinear Function Approximation

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, Le Song

When function approximation is used, solving the Bellman optimality equation with stability guarantees has remained a major open problem in reinforcement learning for decades. The fundamental difficulty is that the Bellman operator may become an expansion in general, resulting in oscillating and even divergent behavior of popular algorithms like Q-learning. In this paper, we revisit the Bellman equation, and reformulate it into a novel primal-dual optimization problem using Nesterov’s smoothing technique and the Legendre-Fenchel transformation. We then develop a new algorithm, called Smoothed Bellman Error Embedding, to solve this optimization problem where any differentiable function class may be used. We provide what we believe to be the first convergence guarantee for general nonlinear function approximation, and analyze the algorithm’s sample complexity. Empirically, our algorithm compares favorably to state-of-the-art baselines in several benchmark control problems.

****

Compressing Neural Networks using the Variational Information Bottleneck

Bin Dai, Chen Zhu, Baining Guo, David Wipf

Neural networks can be compressed to reduce memory and computational requirements, or to increase accuracy by facilitating the use of a larger base architecture. In this paper we focus on pruning individual neurons, which can simultaneously trim model size, FLOPs, and run-time memory. To improve upon the performance of existing compression algorithms we utilize the information bottleneck principle instantiated via a tractable variational bound. Minimization of this information theoretic bound reduces the redundancy between adjacent layers by aggregating useful information into a subset of neurons that can be preserved. In contrast, the activations of disposable neurons are shut off via an attractive form of sparse regularization that emerges naturally from this framework, providing tangible advantages over traditional sparsity penalties without contributing additional tuning parameters to the energy landscape. We demonstrate state-of-the-art compression rates across an array of datasets and network architectures.

****

Asynchronous Byzantine Machine Learning (the case of SGD)

Georgios Damaskinos, El Mahdi El Mhamdi, Rachid Guerraoui, Rhicheek Patra, Mahsa Taziki

Asynchronous distributed machine learning solutions have proven very effective so far, but always assuming perfectly functioning workers. In practice, some of the workers can however exhibit Byzantine behavior, caused by hardware failures, software bugs, corrupt data, or even malicious attacks. We introduce Kardam, the first distributed asynchronous stochastic gradient descent (SGD) algorithm that copes with Byzantine workers. Kardam consists of two complementary components: a filtering and a dampening component. The first is scalar-based and ensures resilience against 1/3 Byzantine workers. Essentially, this filter leverages the Lipschitzness of cost functions and acts as a self-stabilizer against Byzantine workers that would attempt to corrupt the progress of SGD. The dampening component bounds the convergence rate by adjusting to stale information through a generic gradient weighting scheme. We prove that Kardam guarantees almost sure convergence in the presence of asynchrony and Byzantine behavior, and we derive its convergence rate. We evaluate Kardam on the CIFAR100 and EMNIST datasets and measure its overhead with respect to non Byzantine-resilient solutions. We empirically show that Kardam does not introduce additional noise to the learning procedure but does induce a slowdown (the cost of Byzantine resilience) that we both theoretically and empirically show to be less than f/n, where f is the number of Byzantine failures tolerated and n the total number of workers. Interestingly, we also empirically observe that the dampening component is interesting in its own right for it enables to build an SGD algorithm that outperforms alternative staleness-aware asynchronous competitors in environments with honest workers.

****

Escaping Saddles with Stochastic Gradients

Hadi Daneshmand, Jonas Kohler, Aurelien Lucchi, Thomas Hofmann

We analyze the variance of stochastic gradients along negative curvature directions in certain non-convex machine learning models and show that stochastic gradients indeed exhibit a strong component along these directions. Furthermore, we show that - contrary to the case of isotropic noise - this variance is proportional to the magnitude of the corresponding eigenvalues and not decreasing in the dimensionality. Based upon this bservation we propose a new assumption under which we show that the injection of explicit, isotropic noise usually applied to make gradient descent escape saddle points can successfully be replaced by a simple SGD step. Additionally - and under the same condition - we derive the first convergence rate for plain SGD to a second-order stationary point in a number of iterations that is independent of the problem dimension.

****

Minibatch Gibbs Sampling on Large Graphical Models

Chris De Sa, Vincent Chen, Wing Wong

Gibbs sampling is the de facto Markov chain Monte Carlo method used for inference and learning on large scale graphical models. For complicated factor graphs with lots of factors, the performance of Gibbs sampling can be limited by the computational cost of executing a single update step of the Markov chain. This cost is proportional to the degree of the graph, the number of factors adjacent to each variable. In this paper, we show how this cost can be reduced by using minibatching: subsampling the factors to form an estimate of their sum. We introduce several minibatched variants of Gibbs, show that they can be made unbiased, prove bounds on their convergence rates, and show that under some conditions they can result in asymptotic single-update-run-time speedups over plain Gibbs sampling.

****

Stochastic Video Generation with a Learned Prior

Emily Denton, Rob Fergus

Generating video frames that accurately predict future world states is challenging. Existing approaches either fail to capture the full distribution of outcomes, or yield blurry generations, or both. In this paper we introduce a video generation model with a learned prior over stochastic latent variables at each time step. Video frames are generated by drawing samples from this prior and combining them with a deterministic estimate of the future frame. The approach is simple

and easily trained end-to-end on a variety of datasets. Sample generations are b
oth varied and sharp, even many frames into the future, and compare favorably to
 those from existing approaches.
****

Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-se
nsitive Learning
Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, Steffen Udluft
Bayesian neural networks with latent variables are scalable and flexible probabi
listic models: they account for uncertainty in the estimation of the network wei
ghts and, by making use of latent variables, can capture complex noise patterns
in the data. Using these models we show how to perform and utilize a decompositi
on of uncertainty in aleatoric and epistemic components for decision making purp
oses. This allows us to successfully identify informative points for active lear
ning of functions with heteroscedastic and bimodal noise. Using the decompositio
n we further define a novel risk-sensitive criterion for reinforcement learningt
o identify policies that balance expected cost, model-bias and noise aversion.
****

Accurate Inference for Adaptive Linear Models
Yash Deshpande, Lester Mackey, Vasilis Syrgkanis, Matt Taddy
Estimators computed from adaptively collected data do not behave like their non-
adaptive brethren.Rather, the sequential dependence of the collection policy can
 lead to severe distributional biases that persist even in the infinite data lim
it. We develop a general method – $\mathbf{W}$-decorrelation – for transforming
the bias of adaptive linear regression estimators into variance. The method uses
 only coarse-grained information about the data collection policy and does not n
eed access to propensity scores or exact knowledge of the policy.We bound the fi
nite-sample bias and variance of the $\mathbf{W}$-estimator and develop asymptot
ically correct confidence intervals based on a novel martingale central limit th
eorem. We then demonstrate the empirical benefits of the generic $\mathbf{W}$-de
correlation procedure in two different adaptive data settings: the multi-armed b
andit and the autoregressive time series.
****

Variational Network Inference: Strong and Stable with Concrete Support
Amir Dezfouli, Edwin Bonilla, Richard Nock
Traditional methods for the discovery of latent network structures are limited i
n two ways: they either assume that all the signal comes from the network (i.e.
there is no source of signal outside the network) or they place constraints on t
he network parameters to ensure model or algorithmic stability. We address these
 limitations by proposing a model that incorporates a Gaussian process prior on
a network-independent component and formally proving that we get algorithmic sta
bility for free while providing a novel perspective on model stability as well a
s robustness results and precise intervals for key inference parameters. We show
 that, on three applications, our approach outperforms previous methods consiste
ntly.
****

Modeling Sparse Deviations for Compressed Sensing using Generative Models
Manik Dhar, Aditya Grover, Stefano Ermon
In compressed sensing, a small number of linear measurements can be used to reco
nstruct an unknown signal. Existing approaches leverage assumptions on the struc
ture of these signals, such as sparsity or the availability of a generative mode
l. A domain-specific generative model can provide a stronger prior and thus allo
w for recovery with far fewer measurements. However, unlike sparsity-based appro
aches, existing methods based on generative models guarantee exact recovery only
 over their support, which is typically only a small subset of the space on whic
h the signals are defined. We propose Sparse-Gen, a framework that allows for sp
arse deviations from the support set, thereby achieving the best of both worlds
by using a domain specific prior and allowing reconstruction over the full space
 of signals. Theoretically, our framework provides a new class of signals that c
an be acquired using compressed sensing, reducing classic sparse vector recovery
 to a special case and avoiding the restrictive support due to a generative mode

l prior. Empirically, we observe consistent improvements in reconstruction accuracy over competing approaches, especially in the more practical setting of transfer compressed sensing where a generative model for a data-rich, source domain aids sensing on a data-scarce, target domain.
****
Alternating Randomized Block Coordinate Descent
Jelena Diakonikolas, Lorenzo Orecchia
Block-coordinate descent algorithms and alternating minimization methods are fundamental optimization algorithms and an important primitive in large-scale optimization and machine learning. While various block-coordinate-descent-type methods have been studied extensively, only alternating minimization – which applies to the setting of only two blocks – is known to have convergence time that scales independently of the least smooth block. A natural question is then: is the setting of two blocks special? We show that the answer is “no” as long as the least smooth block can be optimized exactly – an assumption that is also needed in the setting of alternating minimization. We do so by introducing a novel algorithm AR-BCD, whose convergence time scales independently of the least smooth (possibly non-smooth) block. The basic algorithm generalizes both alternating minimization and randomized block coordinate (gradient) descent, and we also provide its accelerated version – AAR-BCD.
****
Learning to Act in Decentralized Partially Observable MDPs
Jilles Dibangoye, Olivier Buffet
We address a long-standing open problem of reinforcement learning in decentralized partially observable Markov decision processes. Previous attempts focussed on different forms of generalized policy iteration, which at best led to local optima. In this paper, we restrict attention to plans, which are simpler to store and update than policies. We derive, under certain conditions, the first near-optimal cooperative multi-agent reinforcement learning algorithm. To achieve significant scalability gains, we replace the greedy maximization by mixed-integer linear programming. Experiments show our approach can learn to act near-optimally in many finite domains from the literature.
****
Leveraging Well-Conditioned Bases: Streaming and Distributed Summaries in Minkowski p-Norms
Charlie Dickens, Graham Cormode, David Woodruff
Work on approximate linear algebra has led to efficient distributed and streaming algorithms for problems such as approximate matrix multiplication, low rank approximation, and regression, primarily for the Euclidean norm ℓ_2. We study other ℓ_p norms, which are more robust for $p < 2$, and can be used to find outliers for $p > 2$. Unlike previous algorithms for such norms, we give algorithms that are (1) deterministic, (2) work simultaneously for every $p \geq 1$, including $p = \infty$, and (3) can be implemented in both distributed and streaming environments. We study ℓ_p-regression, entrywise ℓ_p-low rank approximation, and versions of approximate matrix multiplication.
****
Noisin: Unbiased Regularization for Recurrent Neural Networks
Adji Bousso Dieng, Rajesh Ranganath, Jaan Altosaar, David Blei
Recurrent neural networks (RNNs) are powerful models of sequential data. They have been successfully used in domains such as text and speech. However, RNNs are susceptible to overfitting; regularization is important. In this paper we develop Noisin, a new method for regularizing RNNs. Noisin injects random noise into the hidden states of the RNN and then maximizes the corresponding marginal likelihood of the data. We show how Noisin applies to any RNN and we study many different types of noise. Noisin is unbiased–it preserves the underlying RNN on average. We characterize how Noisin regularizes its RNN both theoretically and empirically. On language modeling benchmarks, Noisin improves over dropout by as much as 12.2% on the Penn Treebank and 9.4% on the Wikitext-2 dataset. We also compared the state-of-the-art language model of Yang et al. 2017, both with and without Noisin. On the Penn Treebank, the method with Noisin more quickly reaches state

-of-the-art performance.
****
Discovering and Removing Exogenous State Variables and Rewards for Reinforcement Learning
Thomas Dietterich, George Trimponias, Zhitang Chen
Exogenous state variables and rewards can slow down reinforcement learning by injecting uncontrolled variation into the reward signal. We formalize exogenous state variables and rewards and identify conditions under which an MDP with exogenous state can be decomposed into an exogenous Markov Reward Process involving only the exogenous state+reward and an endogenous Markov Decision Process defined with respect to only the endogenous rewards. We also derive a variance-covariance condition under which Monte Carlo policy evaluation on the endogenous MDP is accelerated compared to using the full MDP. Similar speedups are likely to carry over to all RL algorithms. We develop two algorithms for discovering the exogenous variables and test them on several MDPs. Results show that the algorithms are practical and can significantly speed up reinforcement learning.
****
Coordinated Exploration in Concurrent Reinforcement Learning
Maria Dimakopoulou, Benjamin Van Roy
We consider a team of reinforcement learning agents that concurrently learn to operate in a common environment. We identify three properties - adaptivity, commitment, and diversity - which are necessary for efficient coordinated exploration and demonstrate that straightforward extensions to single-agent optimistic and posterior sampling approaches fail to satisfy them. As an alternative, we propose seed sampling, which extends posterior sampling in a manner that meets these requirements. Simulation results investigate how per-agent regret decreases as the number of agents grows, establishing substantial advantages of seed sampling over alternative exploration schemes.
****
Probabilistic Recurrent State-Space Models
Andreas Doerr, Christian Daniel, Martin Schiegg, Nguyen-Tuong Duy, Stefan Schaal, Marc Toussaint, Trimpe Sebastian
State-space models (SSMs) are a highly expressive model class for learning patterns in time series data and for system identification. Deterministic versions of SSMs (e.g., LSTMs) proved extremely successful in modeling complex time series data. Fully probabilistic SSMs, however, are often found hard to train, even for smaller problems. We propose a novel model formulation and a scalable training algorithm based on doubly stochastic variational inference and Gaussian processes. This combination allows efficient incorporation of latent state temporal correlations, which we found to be key to robust training. The effectiveness of the proposed PR-SSM is evaluated on a set of real-world benchmark datasets in comparison to state-of-the-art probabilistic model learning methods. Scalability and robustness are demonstrated on a high dimensional problem.
****
Randomized Block Cubic Newton Method
Nikita Doikov, Peter Richtarik, University Edinburgh
We study the problem of minimizing the sum of three convex functions: a differentiable, twice-differentiable and a non-smooth term in a high dimensional setting. To this effect we propose and analyze a randomized block cubic Newton (RBCN) method, which in each iteration builds a model of the objective function formed as the sum of the natural models of its three components: a linear model with a quadratic regularizer for the differentiable term, a quadratic model with a cubic regularizer for the twice differentiable term, and perfect (proximal) model for the nonsmooth term. Our method in each iteration minimizes the model over a random subset of blocks of the search variable. RBCN is the first algorithm with these properties, generalizing several existing methods, matching the best known bounds in all special cases. We establish ${\cal O}(1/\epsilon)$, ${\cal O}(1/\sqrt{\epsilon})$ and ${\cal O}(\log (1/\epsilon))$ rates under different assumptions on the component functions. Lastly, we show numerically that our method outperforms the state-of-the-art on a variety of machine learning problems, including

cubically regularized least-squares, logistic regression with constraints, and Poisson regression.
****

Low-Rank Riemannian Optimization on Positive Semidefinite Stochastic Matrices with Applications to Graph Clustering
Ahmed Douik, Babak Hassibi
This paper develops a Riemannian optimization framework for solving optimization problems on the set of symmetric positive semidefinite stochastic matrices. The paper first reformulates the problem by factorizing the optimization variable as $\mathbf{X}=\mathbf{Y}\mathbf{Y}^T$ and deriving conditions on p, i.e., the number of columns of $\mathbf{Y}$, under which the factorization yields a satisfactory solution. The reparameterization of the problem allows its formulation as an optimization over either an embedded or quotient Riemannian manifold whose geometries are investigated. In particular, the paper explicitly derives the tangent space, Riemannian gradients and retraction operator that allow the design of efficient optimization methods on the proposed manifolds. The numerical results reveal that, when the optimal solution has a known low-rank, the resulting algorithms present a clear complexity advantage when compared with state-of-the-art Euclidean and Riemannian approaches for graph clustering applications.
****

Essentially No Barriers in Neural Network Energy Landscape
Felix Draxler, Kambis Veschgini, Manfred Salmhofer, Fred Hamprecht
Training neural networks involves finding minima of a high-dimensional non-convex loss function. Relaxing from linear interpolations, we construct continuous paths between minima of recent neural network architectures on CIFAR10 and CIFAR100. Surprisingly, the paths are essentially flat in both the training and test landscapes. This implies that minima are perhaps best seen as points on a single connected manifold of low loss, rather than as the bottoms of distinct valleys.
****

Weakly Consistent Optimal Pricing Algorithms in Repeated Posted-Price Auctions with Strategic Buyer
Alexey Drutsa
We study revenue optimization learning algorithms for repeated posted-price auctions where a seller interacts with a single strategic buyer that holds a fixed private valuation for a good and seeks to maximize his cumulative discounted surplus. We propose a novel algorithm that never decreases offered prices and has a tight strategic regret bound of $\Theta(\log\log T)$. This result closes the open research question on the existence of a no-regret horizon-independent weakly consistent pricing. We also show that the property of non-decreasing prices is nearly necessary for a weakly consistent algorithm to be a no-regret one.
****

On the Power of Over-parametrization in Neural Networks with Quadratic Activation
Simon Du, Jason Lee
We provide new theoretical insights on why over-parametrization is effective in learning neural networks. For a k hidden node shallow network with quadratic activation and n training data points, we show as long as $ k \ge \sqrt{2n}$, over-parametrization enables local search algorithms to find a globally optimal solution for general smooth and convex loss functions. Further, despite that the number of parameters may exceed the sample size, using theory of Rademacher complexity, we show with weight decay, the solution also generalizes well if the data is sampled from a regular distribution such as Gaussian. To prove when $k\ge \sqrt{2n}$, the loss function has benign landscape properties, we adopt an idea from smoothed analysis, which may have other applications in studying loss surfaces of neural networks.
****

Gradient Descent Learns One-hidden-layer CNN: Don’t be Afraid of Spurious Local Minima
Simon Du, Jason Lee, Yuandong Tian, Aarti Singh, Barnabas Poczos
We consider the problem of learning an one-hidden-layer neural network with non-

overlapping convolutional layer and ReLU activation function, i.e., $f(Z; w, a) = \sum_j a_j\sigma(w^\top Z_j)$, in which both the convolutional weights w and the output weights a are parameters to be learned. We prove that with Gaussian input $\mathbf{Z}$ there is a spurious local minimizer. Surprisingly, in the presence of the spurious local minimizer, starting from randomly initialized weights, gradient descent with weight normalization can still be proven to recover the true parameters with constant probability (which can be boosted to probability 1 with multiple restarts). We also show that with constant probability, the same procedure could also converge to the spurious local minimum, showing that the local minimum plays a non-trivial role in the dynamics of gradient descent. Furthermore, a quantitative analysis shows that the gradient descent dynamics has two phases: it starts off slow, but converges much faster after several iterations.

****

Investigating Human Priors for Playing Video Games

Rachit Dubey, Pulkit Agrawal, Deepak Pathak, Tom Griffiths, Alexei Efros

What makes humans so good at solving seemingly complex video games? Unlike computers, humans bring in a great deal of prior knowledge about the world, enabling efficient decision making. This paper investigates the role of human priors for solving video games. Given a sample game, we conduct a series of ablation studies to quantify the importance of various priors on human performance. We do this by modifying the video game environment to systematically mask different types of visual information that could be used by humans as priors. We find that removal of some prior knowledge causes a drastic degradation in the speed with which human players solve the game, e.g. from 2 minutes to over 20 minutes. Furthermore, our results indicate that general priors, such as the importance of objects and visual consistency, are critical for efficient game-play. Videos and the game manipulations are available at https://rach0012.github.io/humanRL_website/

****

A Distributed Second-Order Algorithm You Can Trust

Celestine Duenner, Aurelien Lucchi, Matilde Gargiani, An Bian, Thomas Hofmann, Martin Jaggi

Due to the rapid growth of data and computational resources, distributed optimization has become an active research area in recent years. While first-order methods seem to dominate the field, second-order methods are nevertheless attractive as they potentially require fewer communication rounds to converge. However, there are significant drawbacks that impede their wide adoption, such as the computation and the communication of a large Hessian matrix. In this paper we present a new algorithm for distributed training of generalized linear models that only requires the computation of diagonal blocks of the Hessian matrix on the individual workers. To deal with this approximate information we propose an adaptive approach that - akin to trust-region methods - dynamically adapts the auxiliary model to compensate for modeling errors. We provide theoretical rates of convergence for a wide class of problems including L_1-regularized objectives. We also demonstrate that our approach achieves state-of-the-art results on multiple large benchmark datasets.

****

Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn’s Algorithm

Pavel Dvurechensky, Alexander Gasnikov, Alexey Kroshnin

We analyze two algorithms for approximating the general optimal transport (OT) distance between two discrete distributions of size n, up to accuracy ε. For the first algorithm, which is based on the celebrated Sinkhorn’s algorithm, we prove the complexity bound $\widetilde{O}\left(\frac{n^2}{\varepsilon^2}\right)$ arithmetic operations ($\widetilde{O}$ hides polylogarithmic factors $(\ln n)^c$, $c>0$). For the second one, which is based on our novel Adaptive Primal-Dual Accelerated Gradient Descent (APDAGD) algorithm, we prove the complexity bound $\widetilde{O}\left(\min\left\{\frac{n^{9/4}}{\varepsilon}, \frac{n^{2}}{\varepsilon^2} \right\}\right)$ arithmetic operations. Both bounds have better dependence on ε than the state-of-the-art result given by $\widetild

e{O}\left(\frac{n^2}{\varepsilon^3}\right)$. Our second algorithm not only has b
etter dependence on ε in the complexity bound, but also is not speci
fic to entropic regularization and can solve the OT problem with different regul
arizers.
****

Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties
of Entropy-SGD and data-dependent priors
Gintare Karolina Dziugaite, Daniel Roy
We show that Entropy-SGD (Chaudhari et al., 2017), when viewed as a learning alg
orithm, optimizes a PAC-Bayes bound on the risk of a Gibbs (posterior) classifie
r, i.e., a randomized classifier obtained by a risk-sensitive perturbation of th
e weights of a learned classifier. Entropy-SGD works by optimizing the bound’s p
rior, violating the hypothesis of the PAC-Bayes theorem that the prior is chosen
 independently of the data. Indeed, available implementations of Entropy-SGD rap
idly obtain zero training error on random labels and the same holds of the Gibbs
 posterior. In order to obtain a valid generalization bound, we rely on a result
 showing that data-dependent priors obtained by stochastic gradient Langevin dyn
amics (SGLD) yield valid PAC-Bayes bounds provided the target distribution of SG
LD is eps-differentially private. We observe that test error on MNIST and CIFAR1
0 falls within the (empirically nonvacuous) risk bounds computed under the assum
ption that SGLD reaches stationarity. In particular, Entropy-SGLD can be configu
red to yield relatively tight generalization bounds and still fit real labels, a
lthough these same settings do not obtain state-of-the-art performance.
****

Beyond the One-Step Greedy Approach in Reinforcement Learning
Yonathan Efroni, Gal Dalal, Bruno Scherrer, Shie Mannor
The famous Policy Iteration algorithm alternates between policy improvement and
policy evaluation. Implementations of this algorithm with several variants of th
e latter evaluation stage, e.g, n-step and trace-based returns, have been analyz
ed in previous works. However, the case of multiple-step lookahead policy improv
ement, despite the recent increase in empirical evidence of its strength, has to
 our knowledge not been carefully analyzed yet. In this work, we introduce the f
irst such analysis. Namely, we formulate variants of multiple-step policy improv
ement, derive new algorithms using these definitions and prove their convergence
. Moreover, we show that recent prominent Reinforcement Learning algorithms are,
 in fact, instances of our framework. We thus shed light on their empirical succ
ess and give a recipe for deriving new algorithms for future study.
****

Parallel and Streaming Algorithms for K-Core Decomposition
Hossein Esfandiari, Silvio Lattanzi, Vahab Mirrokni
The k-core decomposition is a fundamental primitive in many machine learning and
 data mining applications. We present the first distributed and the first stream
ing algorithms to compute and maintain an approximate k-core decomposition with
provable guarantees. Our algorithms achieve rigorous bounds on space complexity
while bounding the number of passes or number of rounds of computation. We do so
 by presenting a new powerful sketching technique for k-core decomposition, and
then by showing it can be computed efficiently in both streaming and MapReduce m
odels. Finally, we confirm the effectiveness of our sketching technique empirica
lly on a number of publicly available graphs.
****

IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Arch
itectures
Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Y
otam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, Koray Kavukcuoglu
In this work we aim to solve a large collection of tasks using a single reinforc
ement learning agent with a single set of parameters. A key challenge is to hand
le the increased amount of data and extended training time. We have developed a
new distributed agent IMPALA (Importance Weighted Actor-Learner Architecture) th
at not only uses resources more efficiently in single-machine training but also
scales to thousands of machines without sacrificing data efficiency or resource

utilisation. We achieve stable learning at high throughput by combining decouple
d acting and learning with a novel off-policy correction method called V-trace.
We demonstrate the effectiveness of IMPALA for multi-task reinforcement learning
 on DMLab-30 (a set of 30 tasks from the DeepMind Lab environment (Beattie et al
., 2016)) and Atari57 (all available Atari games in Arcade Learning Environment
(Bellemare et al., 2013a)). Our results show that IMPALA is able to achieve bett
er performance than previous agents with less data, and crucially exhibits posit
ive transfer between tasks as a result of its multi-task approach.
****

Scalable Gaussian Processes with Grid-Structured Eigenfunctions (GP-GRIEF)
Trefor Evans, Prasanth Nair
We introduce a kernel approximation strategy that enables computation of the Gau
ssian process log marginal likelihood and all hyperparameter derivatives in O(p)
 time. Our GRIEF kernel consists of p eigenfunctions found using a Nystrom appro
ximation from a dense Cartesian product grid of inducing points. By exploiting a
lgebraic properties of Kronecker and Khatri-Rao tensor products, computational c
omplexity of the training procedure can be practically independent of the number
 of inducing points. This allows us to use arbitrarily many inducing points to a
chieve a globally accurate kernel approximation, even in high-dimensional proble
ms. The fast likelihood evaluation enables type-I or II Bayesian inference on la
rge-scale datasets. We benchmark our algorithms on real-world problems with up t
o two-million training points and 10^33 inducing points.
****

The Limits of Maxing, Ranking, and Preference Learning
Moein Falahatgar, Ayush Jain, Alon Orlitsky, Venkatadheeraj Pichapati, Vaishakh
Ravindrakumar
We present a comprehensive understanding of three important problems in PAC pref
erence learning: maximum selection (maxing), ranking, and estimating all pairwis
e preference probabilities, in the adaptive setting. With just Weak Stochastic T
ransitivity, we show that maxing requires $\Omega(n^2)$ comparisons and with sli
ghtly more restrictive Medium Stochastic Transitivity, we present a linear compl
exity maxing algorithm. With Strong Stochastic Transitivity and Stochastic Trian
gle Inequality, we derive a ranking algorithm with optimal $\mathcal{O}(n\log n)
$ complexity and an optimal algorithm that estimates all pairwise preference pro
babilities.
****

BOHB: Robust and Efficient Hyperparameter Optimization at Scale
Stefan Falkner, Aaron Klein, Frank Hutter
Modern deep learning methods are very sensitive to many hyperparameters, and, du
e to the long training times of state-of-the-art models, vanilla Bayesian hyperp
arameter optimization is typically computationally infeasible. On the other hand
, bandit-based configuration evaluation approaches based on random search lack g
uidance and do not converge to the best configurations as quickly. Here, we prop
ose to combine the benefits of both Bayesian optimization and bandit-based metho
ds, in order to achieve the best of both worlds: strong anytime performance and
fast convergence to optimal configurations. We propose a new practical state-of-
the-art hyperparameter optimization method, which consistently outperforms both
Bayesian optimization and Hyperband on a wide range of problem types, including
high-dimensional toy functions, support vector machines, feed-forward neural net
works, Bayesian neural networks, deep reinforcement learning, and convolutional
neural networks. Our method is robust and versatile, while at the same time bein
g conceptually simple and easy to implement.
****

More Robust Doubly Robust Off-policy Evaluation
Mehrdad Farajtabar, Yinlam Chow, Mohammad Ghavamzadeh
We study the problem of off-policy evaluation (OPE) in reinforcement learning (R
L), where the goal is to estimate the performance of a policy from the data gene
rated by another policy(ies). In particular, we focus on the doubly robust (DR)
estimators that consist of an importance sampling (IS) component and a performan
ce model, and utilize the low (or zero) bias of IS and low variance of the model

at the same time. Although the accuracy of the model has a huge impact on the overall performance of DR, most of the work on using the DR estimators in OPE has been focused on improving the IS part, and not much on how to learn the model. In this paper, we propose alternative DR estimators, called more robust doubly robust (MRDR), that learn the model parameter by minimizing the variance of the DR estimator. We first present a formulation for learning the DR model in RL. We then derive formulas for the variance of the DR estimator in both contextual bandits and RL, such that their gradients w.r.t. the model parameters can be estimated from the samples, and propose methods to efficiently minimize the variance. We prove that the MRDR estimators are strongly consistent and asymptotically optimal. Finally, we evaluate MRDR in bandits and RL benchmark problems, and compare its performance with the existing methods.

****

Efficient and Consistent Adversarial Bipartite Matching

Rizal Fathony, Sima Behpour, Xinhua Zhang, Brian Ziebart

Many important structured prediction problems, including learning to rank items, correspondence-based natural language processing, and multi-object tracking, can be formulated as weighted bipartite matching optimizations. Existing structured prediction approaches have significant drawbacks when applied under the constraints of perfect bipartite matchings. Exponential family probabilistic models, such as the conditional random field (CRF), provide statistical consistency guarantees, but suffer computationally from the need to compute the normalization term of its distribution over matchings, which is a #P-hard matrix permanent computation. In contrast, the structured support vector machine (SSVM) provides computational efficiency, but lacks Fisher consistency, meaning that there are distributions of data for which it cannot learn the optimal matching even under ideal learning conditions (i.e., given the true distribution and selecting from all measurable potential functions). We propose adversarial bipartite matching to avoid both of these limitations. We develop this approach algorithmically, establish its computational efficiency and Fisher consistency properties, and apply it to matching problems that demonstrate its empirical benefits.

****

Global Convergence of Policy Gradient Methods for the Linear Quadratic Regulator

Maryam Fazel, Rong Ge, Sham Kakade, Mehran Mesbahi

Direct policy gradient methods for reinforcement learning and continuous control problems are a popular approach for a variety of reasons: 1) they are easy to implement without explicit knowledge of the underlying model, 2) they are an “end-to-end” approach, directly optimizing the performance metric of interest, 3) they inherently allow for richly parameterized policies. A notable drawback is that even in the most basic continuous control problem (that of linear quadratic regulators), these methods must solve a non-convex optimization problem, where little is understood about their efficiency from both computational and statistical perspectives. In contrast, system identification and model based planning in optimal control theory have a much more solid theoretical footing, where much is known with regards to their computational and statistical properties. This work bridges this gap showing that (model free) policy gradient methods globally converge to the optimal solution and are efficient (polynomially so in relevant problem dependent quantities) with regards to their sample and computational complexities.

****

CRVI: Convex Relaxation for Variational Inference

Ghazal Fazelnia, John Paisley

We present a new technique for solving non-convex variational inference optimization problems. Variational inference is a widely used method for posterior approximation in which the inference problem is transformed into an optimization problem. For most models, this optimization is highly non-convex and so hard to solve. In this paper, we introduce a new approach to solving the variational inference optimization based on convex relaxation and semidefinite programming. Our theoretical results guarantee very tight relaxation bounds that get nearer to the global optimal solution than traditional coordinate ascent. We evaluate the perfo

rmance of our approach on regression and sparse coding.
****

Fourier Policy Gradients
Matthew Fellows, Kamil Ciosek, Shimon Whiteson
We propose a new way of deriving policy gradient updates for reinforcement learn ing. Our technique, based on Fourier analysis, recasts integrals that arise with expected policy gradients as convolutions and turns them into multiplications. The obtained analytical solutions allow us to capture the low variance benefits of EPG in a broad range of settings. For the critic, we treat trigonometric and radial basis functions, two function families with the universal approximation p roperty. The choice of policy can be almost arbitrary, including mixtures or hyb rid continuous-discrete probability distributions. Moreover, we derive a general family of sample-based estimators for stochastic policy gradients, which unifie s existing results on sample-based approximation. We believe that this technique has the potential to shape the next generation of policy gradient approaches, p owered by analytical results.
****

Nonparametric variable importance using an augmented neural network with multi-t ask learning
Jean Feng, Brian Williamson, Noah Simon, Marco Carone
In predictive modeling applications, it is often of interest to determine the re lative contribution of subsets of features in explaining the variability of an o utcome. It is useful to consider this variable importance as a function of the u nknown, underlying data-generating mechanism rather than the specific predictive algorithm used to fit the data. In this paper, we connect these ideas in nonpar ametric variable importance to machine learning, and provide a method for effici ent estimation of variable importance when building a predictive model using a n eural network. We show how a single augmented neural network with multi-task lea rning simultaneously estimates the importance of many feature subsets, improving on previous procedures for estimating importance. We demonstrate on simulated d ata that our method is both accurate and computationally efficient, and apply ou r method to both a study of heart disease and for predicting mortality in ICU pa tients.
****

Closed-form Marginal Likelihood in Gamma-Poisson Matrix Factorization
Louis Filstroff, Alberto Lumbreras, Cédric Févotte
We present novel understandings of the Gamma-Poisson (GaP) model, a probabilisti c matrix factorization model for count data. We show that GaP can be rewritten f ree of the score/activation matrix. This gives us new insights about the estimat ion of the topic/dictionary matrix by maximum marginal likelihood estimation. In particular, this explains the robustness of this estimator to over-specified va lues of the factorization rank, especially its ability to automatically prune ir relevant dictionary columns, as empirically observed in previous work. The margi nalization of the activation matrix leads in turn to a new Monte Carlo Expectati on-Maximization algorithm with favorable properties.
****

Automatic Goal Generation for Reinforcement Learning Agents
Carlos Florensa, David Held, Xinyang Geng, Pieter Abbeel
Reinforcement learning (RL) is a powerful technique to train an agent to perform a task; however, an agent that is trained using RL is only capable of achieving the single task that is specified via its reward function. Such an approach doe s not scale well to settings in which an agent needs to perform a diverse set of tasks, such as navigating to varying positions in a room or moving objects to v arying locations. Instead, we propose a method that allows an agent to automatic ally discover the range of tasks that it is capable of performing in its environ ment. We use a generator network to propose tasks for the agent to try to accomp lish, each task being specified as reaching a certain parametrized subset of the state-space. The generator network is optimized using adversarial training to p roduce tasks that are always at the appropriate level of difficulty for the agen t, thus automatically producing a curriculum. We show that, by using this framew

ork, an agent can efficiently and automatically learn to perform a wide set of t
asks without requiring any prior knowledge of its environment, even when only sp
arse rewards are available. Videos and code available at https://sites.google.co
m/view/goalgeneration4rl.
****

DiCE: The Infinitely Differentiable Monte Carlo Estimator

Jakob Foerster, Gregory Farquhar, Maruan Al-Shedivat, Tim Rocktäschel, Eric Xing
, Shimon Whiteson

The score function estimator is widely used for estimating gradients of stochast
ic objectives in stochastic computation graphs (SCG), eg., in reinforcement lear
ning and meta-learning. While deriving the first-order gradient estimators by di
fferentiating a surrogate loss (SL) objective is computationally and conceptuall
y simple, using the same approach for higher-order derivatives is more challengi
ng. Firstly, analytically deriving and implementing such estimators is laborious
 and not compliant with automatic differentiation. Secondly, repeatedly applying
 SL to construct new objectives for each order derivative involves increasingly
cumbersome graph manipulations. Lastly, to match the first-order gradient under
differentiation, SL treats part of the cost as a fixed sample, which we show lea
ds to missing and wrong terms for estimators of higher-order derivatives. To add
ress all these shortcomings in a unified way, we introduce DiCE, which provides
a single objective that can be differentiated repeatedly, generating correct est
imators of derivatives of any order in SCGs. Unlike SL, DiCE relies on automatic
 differentiation for performing the requisite graph manipulations. We verify the
 correctness of DiCE both through a proof and numerical evaluation of the DiCE d
erivative estimates. We also use DiCE to propose and evaluate a novel approach f
or multi-agent learning. Our code is available at https://github.com/alshedivat/
lola.
****

Practical Contextual Bandits with Regression Oracles

Dylan Foster, Alekh Agarwal, Miroslav Dudik, Haipeng Luo, Robert Schapire

A major challenge in contextual bandits is to design general-purpose algorithms
that are both practically useful and theoretically well-founded. We present a ne
w technique that has the empirical and computational advantages of realizability
-based approaches combined with the flexibility of agnostic methods. Our algorit
hms leverage the availability of a regression oracle for the value-function clas
s, a more realistic and reasonable oracle than the classification oracles over p
olicies typically assumed by agnostic methods. Our approach generalizes both UCB
 and LinUCB to far more expressive possible model classes and achieves low regre
t under certain distributional assumptions. In an extensive empirical evaluation
, we find that our approach typically matches or outperforms both realizability-
based and agnostic baselines.
****

Generative Temporal Models with Spatial Memory for Partially Observed Environmen
ts

Marco Fraccaro, Danilo Rezende, Yori Zwols, Alexander Pritzel, S. M. Ali Eslami,
 Fabio Viola

In model-based reinforcement learning, generative and temporal models of environ
ments can be leveraged to boost agent performance, either by tuning the agent’s
representations during training or via use as part of an explicit planning mecha
nism. However, their application in practice has been limited to simplistic envi
ronments, due to the difficulty of training such models in larger, potentially p
artially-observed and 3D environments. In this work we introduce a novel action-
conditioned generative model of such challenging environments. The model feature
s a non-parametric spatial memory system in which we store learned, disentangled
 representations of the environment. Low-dimensional spatial updates are compute
d using a state-space model that makes use of knowledge on the prior dynamics of
 the moving agent, and high-dimensional visual observations are modelled with a
Variational Auto-Encoder. The result is a scalable architecture capable of perfo
rming coherent predictions over hundreds of time steps across a range of partial
ly observed 2D and 3D environments.

****
ADMM and Accelerated ADMM as Continuous Dynamical Systems
Guilherme Franca, Daniel Robinson, Rene Vidal
Recently, there has been an increasing interest in using tools from dynamical sy stems to analyze the behavior of simple optimization algorithms such as gradient descent and accelerated variants. This paper strengthens such connections by de riving the differential equations that model the continuous limit of the sequenc e of iterates generated by the alternating direction method of multipliers, as w ell as an accelerated variant. We employ the direct method of Lyapunov to analyz e the stability of critical points of the dynamical systems and to obtain associ ated convergence rates.
****
Bilevel Programming for Hyperparameter Optimization and Meta-Learning
Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, Massimiliano Po ntil
We introduce a framework based on bilevel programming that unifies gradient-base d hyperparameter optimization and meta-learning. We show that an approximate ver sion of the bilevel problem can be solved by taking into explicit account the op timization dynamics for the inner objective. Depending on the specific setting, the outer variables take either the meaning of hyperparameters in a supervised l earning problem or parameters of a meta-learner. We provide sufficient condition s under which solutions of the approximate problem converge to those of the exac t problem. We instantiate our approach for meta-learning in the case of deep lea rning where representation layers are treated as hyperparameters shared across a set of training episodes. In experiments, we confirm our theoretical findings, present encouraging results for few-shot learning and contrast the bilevel appro ach against classical approaches for learning-to-learn.
****
Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learni ng
Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, Ronald Ortner
We introduce SCAL, an algorithm designed to perform efficient exploration-explor ation in any unknown weakly-communicating Markov Decision Process (MDP) for whic h an upper bound c on the span of the optimal bias function is known. For an MDP with S states, A actions and $\Gamma \leq S$ possible next states, we prove a regret bound of $O(c\sqrt{\Gamma SAT})$, which significantly improves over ex isting algorithms (e.g., UCRL and PSRL), whose regret scales linearly with the M DP diameter D. In fact, the optimal bias span is finite and often much smaller than D (e.g., $D=+\infty$ in non-communicating MDPs). A similar result was or iginally derived by Bartlett and Tewari (2009) for REGAL.C, for which no tractab le algorithm is available. In this paper, we relax the optimization problem at t he core of REGAL.C, we carefully analyze its properties, and we provide the firs t computationally efficient algorithm to solve it. Finally, we report numerical simulations supporting our theoretical findings and showing how SCAL significant ly outperforms UCRL in MDPs with large diameter and small span.
****
Addressing Function Approximation Error in Actor-Critic Methods
Scott Fujimoto, Herke Hoof, David Meger
In value-based reinforcement learning methods such as deep Q-learning, function approximation errors are known to lead to overestimated value estimates and subo ptimal policies. We show that this problem persists in an actor-critic setting a nd propose novel mechanisms to minimize its effects on both the actor and the cr itic. Our algorithm builds on Double Q-learning, by taking the minimum value bet ween a pair of critics to limit overestimation. We draw the connection between t arget networks and overestimation bias, and suggest delaying policy updates to r educe per-update error and further improve performance. We evaluate our method o n the suite of OpenAI gym tasks, outperforming the state of the art in every env ironment tested.
****
Clipped Action Policy Gradient

Yasuhiro Fujita, Shin-ichi Maeda
Many continuous control tasks have bounded action spaces. When policy gradient methods are applied to such tasks, out-of-bound actions need to be clipped before execution, while policies are usually optimized as if the actions are not clipped. We propose a policy gradient estimator that exploits the knowledge of actions being clipped to reduce the variance in estimation. We prove that our estimator, named clipped action policy gradient (CAPG), is unbiased and achieves lower variance than the conventional estimator that ignores action bounds. Experimental results demonstrate that CAPG generally outperforms the conventional estimator, indicating that it is a better policy gradient estimator for continuous control tasks. The source code is available at https://github.com/pfnet-research/capg.
****

Born Again Neural Networks
Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, Anima Anandkumar
Knowledge Distillation (KD) consists of transferring “knowledge” from one machine learning model (the teacher) to another (the student). Commonly, the teacher is a high-capacity model with formidable performance, while the student is more compact. By transferring knowledge, one hopes to benefit from the student’s compactness, without sacrificing too much performance. We study KD from a new perspective: rather than compressing models, we train students parameterized identically to their teachers. Surprisingly, these Born-Again Networks (BANs), outperform their teachers significantly, both on computer vision and language modeling tasks. Our experiments with BANs based on DenseNets demonstrate state-of-the-art performance on the CIFAR-10 (3.5%) and CIFAR-100 (15.5%) datasets, by validation error. Additional experiments explore two distillation objectives: (i) Confidence-Weighted by Teacher Max (CWTM) and (ii) Dark Knowledge with Permuted Predictions (DKPP). Both methods elucidate the essential components of KD, demonstrating the effect of the teacher outputs on both predicted and non-predicted classes.
****

The Generalization Error of Dictionary Learning with Moreau Envelopes
Alexandros Georgogiannis
This is a theoretical study on the sample complexity of dictionary learning with a general type of reconstruction loss. The goal is to estimate a $m \times d$ matrix D of unit-norm columns when the only available information is a set of training samples. Points x in $\mathbb{R}^m$ are subsequently approximated by the linear combination Da after solving the problem $\min_{a \in \mathbb{R}^d} \Phi(x - Da) + g(a)$; function $g:\mathbb{R}^d \to [0,+\infty)$ is either an indicator function or a sparsity promoting regularizer. Here is considered the case where $ \Phi(x) = \inf_{z \in \mathbb{R}^m} { ||x-z||_2^2 + h(||z||_2)}$ and h is an even and univariate function on the real line. Connections are drawn between Φ and the Moreau envelope of h. A new sample complexity result concerning the k-sparse dictionary problem removes the spurious condition on the coherence of D appearing in previous works. Finally, comments are made on the approximation error of certain families of losses. The derived generalization bounds are of order $\mathcal{O}(\sqrt{\log n /n})$ and valid without any further restrictions on the set of dictionaries with unit-norm columns.
****

Local Private Hypothesis Testing: Chi-Square Tests
Marco Gaboardi, Ryan Rogers
The local model for differential privacy is emerging as the reference model for practical applications of collecting and sharing sensitive information while satisfying strong privacy guarantees. In the local model, there is no trusted entity which is allowed to have each individual’s raw data as is assumed in the traditional curator model. Individuals’ data are usually perturbed before sharing them. We explore the design of private hypothesis tests in the local model, where each data entry is perturbed to ensure the privacy of each participant. Specifically, we analyze locally private chi-square tests for goodness of fit and independence testing.
****

Inductive Two-Layer Modeling with Parametric Bregman Transfer
Vignesh Ganapathiraman, Zhan Shi, Xinhua Zhang, Yaoliang Yu
Latent prediction models, exemplified by multi-layer networks, employ hidden variables that automate abstract feature discovery. They typically pose nonconvex optimization problems and effective semi-definite programming (SDP) relaxations have been developed to enable global solutions (Aslan et al., 2014).However, these models rely on nonparametric training of layer-wise kernel representations, and are therefore restricted to transductive learning which slows down test prediction. In this paper, we develop a new inductive learning framework for parametric transfer functions using matching losses. The result for ReLU utilizes completely positive matrices, and the inductive learner not only delivers superior accuracy but also offers an order of magnitude speedup over SDP with constant approximation guarantees.
****

Hyperbolic Entailment Cones for Learning Hierarchical Embeddings
Octavian Ganea, Gary Becigneul, Thomas Hofmann
Learning graph representations via low-dimensional embeddings that preserve relevant network properties is an important class of problems in machine learning. We here present a novel method to embed directed acyclic graphs. Following prior work, we first advocate for using hyperbolic spaces which provably model tree-like structures better than Euclidean geometry. Second, we view hierarchical relations as partial orders defined using a family of nested geodesically convex cones. We prove that these entailment cones admit an optimal shape with a closed form expression both in the Euclidean and hyperbolic spaces, and they canonically define the embedding learning process. Experiments show significant improvements of our method over strong recent baselines both in terms of representational capacity and generalization.
****

Parameterized Algorithms for the Matrix Completion Problem
Robert Ganian, Iyad Kanj, Sebastian Ordyniak, Stefan Szeider
We consider two matrix completion problems, in which we are given a matrix with missing entries and the task is to complete the matrix in a way that (1) minimizes the rank, or (2) minimizes the number of distinct rows. We study the parameterized complexity of the two aforementioned problems with respect to several parameters of interest, including the minimum number of matrix rows, columns, and rows plus columns needed to cover all missing entries. We obtain new algorithmic results showing that, for the bounded domain case, both problems are fixed-parameter tractable with respect to all aforementioned parameters. We complement these results with a lower-bound result for the unbounded domain case that rules out fixed-parameter tractability w.r.t. some of the parameters under consideration.
****

Synthesizing Programs for Images using Reinforced Adversarial Learning
Yaroslav Ganin, Tejas Kulkarni, Igor Babuschkin, S. M. Ali Eslami, Oriol Vinyals
Advances in deep generative networks have led to impressive results in recent years. Nevertheless, such models can often waste their capacity on the minutiae of datasets, presumably due to weak inductive biases in their decoders. This is where graphics engines may come in handy since they abstract away low-level details and represent images as high-level programs. Current methods that combine deep learning and renderers are limited by hand-crafted likelihood or distance functions, a need for large amounts of supervision, or difficulties in scaling their inference algorithms to richer datasets. To mitigate these issues, we present SPIRAL, an adversarially trained agent that generates a program which is executed by a graphics engine to interpret and sample images. The goal of this agent is to fool a discriminator network that distinguishes between real and rendered data, trained with a distributed reinforcement learning setup without any supervision. A surprising finding is that using the discriminator’s output as a reward signal is the key to allow the agent to make meaningful progress at matching the desired output rendering. To the best of our knowledge, this is the first demonstration of an end-to-end, unsupervised and adversarial inverse graphics agent on challenging real world (MNIST, Omniglot, CelebA) and synthetic 3D datasets. A vid

eo of the agent can be found at https://youtu.be/iSyvwAwa7vk.
****
Spotlight: Optimizing Device Placement for Training Deep Neural Networks
Yuanxiang Gao, Li Chen, Baochun Li
Training deep neural networks (DNNs) requires an increasing amount of computation resources, and it becomes typical to use a mixture of GPU and CPU devices. Due to the heterogeneity of these devices, a recent challenge is how each operation in a neural network can be optimally placed on these devices, so that the training process can take the shortest amount of time possible. The current state-of-the-art solution uses reinforcement learning based on the policy gradient method, and it suffers from suboptimal training times. In this paper, we propose Spotlight, a new reinforcement learning algorithm based on proximal policy optimization, designed specifically for finding an optimal device placement for training DNNs. The design of our new algorithm relies upon a new model of the device placement problem: by modeling it as a Markov decision process with multiple stages, we are able to prove that Spotlight achieves a theoretical guarantee on performance improvements. We have implemented Spotlight in the CIFAR-10 benchmark and deployed it on the Google Cloud platform. Extensive experiments have demonstrated that the training time with placements recommended by Spotlight is 60.9% of that recommended by the policy gradient method.
****
Parallel Bayesian Network Structure Learning
Tian Gao, Dennis Wei
Recent advances in Bayesian Network (BN) structure learning have focused on local-to-global learning, where the graph structure is learned via one local subgraph at a time. As a natural progression, we investigate parallel learning of BN structures via multiple learning agents simultaneously, where each agent learns on one local subgraph at a time. We find that parallel learning can reduce the number of subgraphs requiring structure learning by storing previously queried results and communicating (even partial) results among agents. More specifically, by using novel rules on query subset and superset inference, many subgraph structures can be inferred without learning. We provide a sound and complete parallel structure learning (PSL) algorithm, and demonstrate its improved efficiency over state-of-the-art single-thread learning algorithms.
****
Structured Output Learning with Abstention: Application to Accurate Opinion Prediction
Alexandre Garcia, Chloé Clavel, Slim Essid, Florence d’Alche-Buc
Motivated by Supervised Opinion Analysis, we propose a novel framework devoted to Structured Output Learning with Abstention (SOLA). The structure prediction model is able to abstain from predicting some labels in the structured output at a cost chosen by the user in a flexible way. For that purpose, we decompose the problem into the learning of a pair of predictors, one devoted to structured abstention and the other, to structured output prediction. To compare fully labeled training data with predictions potentially containing abstentions, we define a wide class of asymmetric abstention-aware losses. Learning is achieved by surrogate regression in an appropriate feature space while prediction with abstention is performed by solving a new pre-image problem. Thus, SOLA extends recent ideas about Structured Output Prediction via surrogate problems and calibration theory and enjoys statistical guarantees on the resulting excess risk. Instantiated on a hierarchical abstention-aware loss, SOLA is shown to be relevant for fine-grained opinion mining and gives state-of-the-art results on this task. Moreover, the abstention-aware representations can be used to competitively predict user-review ratings based on a sentence-level opinion predictor.
****
Conditional Neural Processes
Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, S. M. Ali Eslami
Deep neural networks excel at function approximation, yet they are typically trained from scratch for each new function. On the other hand, Bayesian methods, su

ch as Gaussian Processes (GPs), exploit prior knowledge to quickly infer the sha
pe of a new function at test time. Yet, GPs are computationally expensive, and i
t can be hard to design appropriate priors. In this paper we propose a family of
 neural models, Conditional Neural Processes (CNPs), that combine the benefits o
f both. CNPs are inspired by the flexibility of stochastic processes such as GPs
, but are structured as neural networks and trained via gradient descent. CNPs m
ake accurate predictions after observing only a handful of training data points,
 yet scale to complex functions and large datasets. We demonstrate the performan
ce and versatility of the approach on a range of canonical machine learning task
s, including regression, classification and image completion.
****

Temporal Poisson Square Root Graphical Models
Sinong Geng, Zhaobin Kuang, Peggy Peissig, David Page
We propose temporal Poisson square root graphical models (TPSQRs), a generalizat
ion of Poisson square root graphical models (PSQRs) specifically designed for mo
deling longitudinal event data. By estimating the temporal relationships for all
 possible pairs of event types, TPSQRs can offer a holistic perspective about wh
ether the occurrences of any given event type could excite or inhibit any other
type. A TPSQR is learned by estimating a collection of interrelated PSQRs that s
hare the same template parameterization. These PSQRs are estimated jointly in a
pseudo-likelihood fashion, where Poisson pseudo-likelihood is used to approximat
e the original more computationally intensive pseudo-likelihood problem stemming
 from PSQRs. Theoretically, we demonstrate that under mild assumptions, the Pois
son pseudolikelihood approximation is sparsistent for recovering the underlying
PSQR. Empirically, we learn TPSQRs from a real-world large-scale electronic heal
th record (EHR) with millions of drug prescription and condition diagnosis event
s, for adverse drug reaction (ADR) detection. Experimental results demonstrate t
hat the learned TPSQRs can recover ADR signals from the EHR effectively and effi
ciently.
****

Budgeted Experiment Design for Causal Structure Learning
AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, Elias Bareinboim
We study the problem of causal structure learning when the experimenter is limit
ed to perform at most k non-adaptive experiments of size 1. We formulate the
 problem of finding the best intervention target set as an optimization problem,
 which aims to maximize the average number of edges whose directions are resolve
d. We prove that the corresponding objective function is submodular and a greedy
 algorithm suffices to achieve $(1-\frac{1}{e})$-approximation of the optimal va
lue. We further present an accelerated variant of the greedy algorithm, which ca
n lead to orders of magnitude performance speedup. We validate our proposed appr
oach on synthetic and real graphs. The results show that compared to the purely
observational setting, our algorithm orients the majority of the edges through a
 considerably small number of interventions.
****

Linear Spectral Estimators and an Application to Phase Retrieval
Ramina Ghods, Andrew Lan, Tom Goldstein, Christoph Studer
Phase retrieval refers to the problem of recovering real- or complex-valued vect
ors from magnitude measurements. The best-known algorithms for this problem are
iterative in nature and rely on so-called spectral initializers that provide acc
urate initialization vectors. We propose a novel class of estimators suitable fo
r general nonlinear measurement systems, called linear spectral estimators (LSPE
s), which can be used to compute accurate initialization vectors for phase retri
eval problems. The proposed LSPEs not only provide accurate initialization vecto
rs for noisy phase retrieval systems with structured or random measurement matri
ces, but also enable the derivation of sharp and nonasymptotic mean-squared erro
r bounds. We demonstrate the efficacy of LSPEs on synthetic and real-world phase
 retrieval problems, and we show that our estimators significantly outperform ex
isting methods for structured measurement systems that arise in practice.
****

Structured Variational Learning of Bayesian Neural Networks with Horseshoe Prior

s
Soumya Ghosh, Jiayu Yao, Finale Doshi-Velez
Bayesian Neural Networks (BNNs) have recently received increasing attention for their ability to provide well-calibrated posterior uncertainties. However, model selection—even choosing the number of nodes—remains an open question. Recent work has proposed the use of a horseshoe prior over node pre-activations of a Bayesian neural network, which effectively turns off nodes that do not help explain the data. In this work, we propose several modeling and inference advances that consistently improve the compactness of the model learned while maintaining predictive performance, especially in smaller-sample settings including reinforcement learning.
****
Learning Maximum-A-Posteriori Perturbation Models for Structured Prediction in Polynomial Time
Asish Ghoshal, Jean Honorio
MAP perturbation models have emerged as a powerful framework for inference in structured prediction. Such models provide a way to efficiently sample from the Gibbs distribution and facilitate predictions that are robust to random noise. In this paper, we propose a provably polynomial time randomized algorithm for learning the parameters of perturbed MAP predictors. Our approach is based on minimizing a novel Rademacher-based generalization bound on the expected loss of a perturbed MAP predictor, which can be computed in polynomial time. We obtain conditions under which our randomized learning algorithm can guarantee generalization to unseen examples.
****
Robust and Scalable Models of Microbiome Dynamics
Travis Gibson, Georg Gerber
Microbes are everywhere, including in and on our bodies, and have been shown to play key roles in a variety of prevalent human diseases. Consequently, there has been intense interest in the design of bacteriotherapies or "bugs as drugs," which are communities of bacteria administered to patients for specific therapeutic applications. Central to the design of such therapeutics is an understanding of the causal microbial interaction network and the population dynamics of the organisms. In this work we present a Bayesian nonparametric model and associated efficient inference algorithm that addresses the key conceptual and practical challenges of learning microbial dynamics from time series microbe abundance data. These challenges include high-dimensional (300+ strains of bacteria in the gut) but temporally sparse and non-uniformly sampled data; high measurement noise; and, nonlinear and physically non-negative dynamics. Our contributions include a new type of dynamical systems model for microbial dynamics based on what we term interaction modules, or learned clusters of latent variables with redundant interaction structure (reducing the expected number of interaction coefficients from O(n^2) to O((log n)^2)); a fully Bayesian formulation of the stochastic dynamical systems model that propagates measurement and latent state uncertainty throughout the model; and introduction of a temporally varying auxiliary variable technique to enable efficient inference by relaxing the hard non-negativity constraint on states. We apply our method to simulated and real data, and demonstrate the utility of our technique for system identification from limited data and gaining new biological insights into bacteriotherapy design.
****
Non-linear motor control by local learning in spiking neural networks
Aditya Gilra, Wulfram Gerstner
Learning weights in a spiking neural network with hidden neurons, using local, stable and online rules, to control non-linear body dynamics is an open problem. Here, we employ a supervised scheme, Feedback-based Online Local Learning Of Weights (FOLLOW), to train a heterogeneous network of spiking neurons with hidden layers, to control a two-link arm so as to reproduce a desired state trajectory. We show that the network learns an inverse model of the non-linear dynamics, i.e. it infers from state trajectory as input to the network, the continuous-time command that produced the trajectory. Connection weights are adjusted via a local

plasticity rule that involves pre-synaptic firing and post-synaptic feedback of the error in the inferred command. We propose a network architecture, termed differential feedforward, and show that it gives a lower test error than other feedforward and recurrent architectures. We demonstrate the performance of the inverse model to control a two-link arm along a desired trajectory.
****

Learning One Convolutional Layer with Overlapping Patches
Surbhi Goel, Adam Klivans, Raghu Meka
We give the first provably efficient algorithm for learning a one hidden layer convolutional network with respect to a general class of (potentially overlapping) patches under mild conditions on the underlying distribution. We prove that our framework captures commonly used schemes from computer vision, including one-dimensional and two-dimensional “patch and stride” convolutions. Our algorithm– Convotron– is inspired by recent work applying isotonic regression to learning neural networks. Convotron uses a simple, iterative update rule that is stochastic in nature and tolerant to noise (requires only that the conditional mean function is a one layer convolutional network, as opposed to the realizable setting). In contrast to gradient descent, Convotron requires no special initialization or learning-rate tuning to converge to the global optimum. We also point out that learning one hidden convolutional layer with respect to a Gaussian distribution and just one disjoint patch P (the other patches may be arbitrary) is easy in the following sense: Convotron can efficiently recover the hidden weight vector by updating only in the direction of P.
****

Visualizing and Understanding Atari Agents
Samuel Greydanus, Anurag Koul, Jonathan Dodge, Alan Fern
While deep reinforcement learning (deep RL) agents are effective at maximizing rewards, it is often unclear what strategies they use to do so. In this paper, we take a step toward explaining deep RL agents through a case study using Atari 2600 environments. In particular, we focus on using saliency maps to understand how an agent learns and executes a policy. We introduce a method for generating useful saliency maps and use it to show 1) what strong agents attend to, 2) whether agents are making decisions for the right or wrong reasons, and 3) how agents evolve during learning. We also test our method on non-expert human subjects and find that it improves their ability to reason about these agents. Overall, our results show that saliency information can provide significant insight into an RL agent’s decisions and learning behavior.
****

Learning Policy Representations in Multiagent Systems
Aditya Grover, Maruan Al-Shedivat, Jayesh Gupta, Yuri Burda, Harrison Edwards
Modeling agent behavior is central to understanding the emergence of complex phenomena in multiagent systems. Prior work in agent modeling has largely been task-specific and driven by hand-engineering domain-specific prior knowledge. We propose a general learning framework for modeling agent behavior in any multiagent system using only a handful of interaction data. Our framework casts agent modeling as a representation learning problem. Consequently, we construct a novel objective inspired by imitation learning and agent identification and design an algorithm for unsupervised learning of representations of agent policies. We demonstrate empirically the utility of the proposed framework in (i) a challenging high-dimensional competitive environment for continuous control and (ii) a cooperative environment for communication, on supervised predictive tasks, unsupervised clustering, and policy optimization using deep reinforcement learning.
****

Faster Derivative-Free Stochastic Algorithm for Shared Memory Machines
Bin Gu, Zhouyuan Huo, Cheng Deng, Heng Huang
Asynchronous parallel stochastic gradient optimization has been playing a pivotal role to solve large-scale machine learning problems in big data applications. Zeroth-order (derivative-free) methods estimate the gradient only by two function evaluations, thus have been applied to solve the problems where the explicit gradient calculations are computationally expensive or infeasible. Recently, the

first asynchronous parallel stochastic zeroth-order algorithm (AsySZO) was propo sed. However, its convergence rate is O(1/SQRT{T}) for the smooth, possibly non-convex learning problems, which is significantly slower than O(1/T) the best con vergence rate of (asynchronous) stochastic gradient algorithm. To fill this gap, in this paper, we first point out the fundamental reason leading to the slow co nvergence rate of AsySZO, and then propose a new asynchronous stochastic zeroth o rder algorithm (AsySZO+). We provide a faster convergence rate O(1/bT) (b is the mini-batch size) for AsySZO+ by the rigorous theoretical analysis, which is a s ignificant improvement over O(1/SQRT{T}). The experimental results on the applic ation of ensemble learning confirm that our AsySZO+ has a faster convergence rat e than the existing (asynchronous) stochastic zeroth-order algorithms.
****

Learning to search with MCTSnets
Arthur Guez, Theophane Weber, Ioannis Antonoglou, Karen Simonyan, Oriol Vinyals, Daan Wierstra, Remi Munos, David Silver
Planning problems are among the most important and well-studied problems in arti ficial intelligence. They are most typically solved by tree search algorithms th at simulate ahead into the future, evaluate future states, and back-up those eva luations to the root of a search tree. Among these algorithms, Monte-Carlo tree search (MCTS) is one of the most general, powerful and widely used. A typical im plementation of MCTS uses cleverly designed rules, optimised to the particular c haracteristics of the domain. These rules control where the simulation traverses , what to evaluate in the states that are reached, and how to back-up those eval uations. In this paper we instead learn where, what and how to search. Our archi tecture, which we call an MCTSnet, incorporates simulation-based search inside a neural network, by expanding, evaluating and backing-up a vector embedding. The parameters of the network are trained end-to-end using gradient-based optimisat ion. When applied to small searches in the well-known planning problem Sokoban, the learned search algorithm significantly outperformed MCTS baselines.
****

Characterizing Implicit Bias in Terms of Optimization Geometry
Suriya Gunasekar, Jason Lee, Daniel Soudry, Nathan Srebro
We study the bias of generic optimization methods, including Mirror Descent, Nat ural Gradient Descent and Steepest Descent with respect to different potentials and norms, when optimizing underdetermined linear models or separable linear cla ssification problems. We ask the question of whether the global minimum (among t he many possible global minima) reached by optimization can be characterized in terms of the potential or norm, and indecently of hyper-parameter choices, such as stepsize and momentum.
****

Shampoo: Preconditioned Stochastic Tensor Optimization
Vineet Gupta, Tomer Koren, Yoram Singer
Preconditioned gradient methods are among the most general and powerful tools in optimization. However, preconditioning requires storing and manipulating prohib itively large matrices. We describe and analyze a new structure-aware preconditi oning algorithm, called Shampoo, for stochastic optimization over tensor spaces. Shampoo maintains a set of preconditioning matrices, each of which operates on a single dimension, contracting over the remaining dimensions. We establish conv ergence guarantees in the stochastic convex setting, the proof of which builds u pon matrix trace inequalities. Our experiments with state-of-the-art deep learni ng models show that Shampoo is capable of converging considerably faster than co mmonly used optimizers. Surprisingly, although it involves a more complex update rule, Shampoo’s runtime per step is comparable in practice to that of simple gr adient methods such as SGD, AdaGrad, and Adam.
****

Latent Space Policies for Hierarchical Reinforcement Learning
Tuomas Haarnoja, Kristian Hartikainen, Pieter Abbeel, Sergey Levine
We address the problem of learning hierarchical deep neural network policies for reinforcement learning. In contrast to methods that explicitly restrict or crip ple lower layers of a hierarchy to force them to use higher-level modulating sig

nals, each layer in our framework is trained to directly solve the task, but acq uires a range of diverse strategies via a maximum entropy reinforcement learning objective. Each layer is also augmented with latent random variables, which are sampled from a prior distribution during the training of that layer. The maximu m entropy objective causes these latent variables to be incorporated into the la yer’s policy, and the higher level layer can directly control the behavior of th e lower layer through this latent space. Furthermore, by constraining the mappin g from latent variables to actions to be invertible, higher layers retain full e xpressivity: neither the higher layers nor the lower layers are constrained in t heir behavior. Our experimental evaluation demonstrates that we can improve on t he performance of single-layer policies on standard benchmark tasks simply by ad ding additional layers, and that our method can solve more complex sparse-reward tasks by learning higher-level policies on top of high-entropy skills optimized for simple low-level objectives.
****

Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor
Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, Sergey Levine
Model-free deep reinforcement learning (RL) algorithms have been demonstrated on a range of challenging decision making and control tasks. However, these method s typically suffer from two major challenges: very high sample complexity and br ittle convergence properties, which necessitate meticulous hyperparameter tuning . Both of these challenges severely limit the applicability of such methods to c omplex, real-world domains. In this paper, we propose soft actor-critic, an off- policy actor-critic deep RL algorithm based on the maximum entropy reinforcement learning framework. In this framework, the actor aims to maximize expected rewa rd while also maximizing entropy. That is, to succeed at the task while acting a s randomly as possible. Prior deep RL methods based on this framework have been formulated as Q-learning methods. By combining off-policy updates with a stable stochastic actor-critic formulation, our method achieves state-of-the-art perfor mance on a range of continuous control benchmark tasks, outperforming prior on-p olicy and off-policy methods. Furthermore, we demonstrate that, in contrast to o ther off-policy algorithms, our approach is very stable, achieving very similar performance across different random seeds.
****

Comparison-Based Random Forests
Siavash Haghiri, Damien Garreau, Ulrike Luxburg
Assume we are given a set of items from a general metric space, but we neither h ave access to the representation of the data nor to the distances between data p oints. Instead, suppose that we can actively choose a triplet of items (A, B, C) and ask an oracle whether item A is closer to item B or to item C. In this pape r, we propose a novel random forest algorithm for regression and classification that relies only on such triplet comparisons. In the theory part of this paper, we establish sufficient conditions for the consistency of such a forest. In a se t of comprehensive experiments, we then demonstrate that the proposed random for est is efficient both for classification and regression. In particular, it is ev en competitive with other methods that have direct access to the metric represen tation of the data.
****

K-Beam Minimax: Efficient Optimization for Deep Adversarial Learning
Jihun Hamm, Yung-Kyun Noh
Minimax optimization plays a key role in adversarial training of machine learnin g algorithms, such as learning generative models, domain adaptation, privacy pre servation, and robust learning. In this paper, we demonstrate the failure of alt ernating gradient descent in minimax optimization problems due to the discontinu ity of solutions of the inner maximization. To address this, we propose a new $\ epsilon$-subgradient descent algorithm that addresses this problem by simultaneo usly tracking K candidate solutions. Practically, the algorithm can find solut ions that previous saddle-point algorithms cannot find, with only a sublinear in crease of complexity in K. We analyze the conditions under which the algorithm

converges to the true solution in detail. A significant improvement in stabilit y and convergence speed of the algorithm is observed in simple representative pr oblems, GAN training, and domain-adaptation problems.
****

Candidates vs. Noises Estimation for Large Multi-Class Classification Problem
Lei Han, Yiheng Huang, Tong Zhang
This paper proposes a method for multi-class classification problems, where the number of classes K is large. The method, referred to as Candidates vs. Noises E stimation (CANE), selects a small subset of candidate classes and samples the re maining classes. We show that CANE is always consistent and computationally effi cient. Moreover, the resulting estimator has low statistical variance approachin g that of the maximum likelihood estimator, when the observed label belongs to t he selected candidates with high probability. In practice, we use a tree structu re with leaves as classes to promote fast beam search for candidate selection. W e further apply the CANE method to estimate word probabilities in learning large neural language models. Extensive experimental results show that CANE achieves better prediction accuracy over the Noise-Contrastive Estimation (NCE), its vari ants and a number of the state-of-the-art tree classifiers, while it gains signi ficant speedup compared to standard O(K) methods.
****

Stein Variational Gradient Descent Without Gradient
Jun Han, Qiang Liu
Stein variational gradient decent (SVGD) has been shown to be a powerful approxi mate inference algorithm for complex distributions. However, the standard SVGD r equires calculating the gradient of the target density and cannot be applied whe n the gradient is unavailable. In this work, we develop a gradient-free variant of SVGD (GF-SVGD), which replaces the true gradient with a surrogate gradient, a nd corrects the introduced bias by re-weighting the gradients in a proper form. We show that our GF-SVGD can be viewed as the standard SVGD with a special choic e of kernel, and hence directly inherits all the theoretical properties of SVGD. We shed insights on the empirical choice of the surrogate gradient and further, propose an annealed GF-SVGD that consistently outperforms a number of recent ad vanced gradient-free MCMC methods in our empirical studies.
****

Deep Models of Interactions Across Sets
Jason Hartford, Devon Graham, Kevin Leyton-Brown, Siamak Ravanbakhsh
We use deep learning to model interactions across two or more sets of objects, s uch as user{–}movie ratings or protein{–}drug bindings. The canonical representa tion of such interactions is a matrix (or tensor) with an exchangeability proper ty: the encoding’s meaning is not changed by permuting rows or columns. We argue that models should hence be Permutation Equivariant (PE): constrained to make t he same predictions across such permutations. We present a parameter-sharing sch eme and prove that it is maximally expressive under the PE constraint. This sche me yields three benefits. First, we demonstrate performance competitive with the state of the art on multiple matrix completion benchmarks. Second, our models r equire a number of parameters independent of the numbers of objects and thus sca le well to large datasets. Third, models can be queried about new objects that w ere not available at training time, but for which interactions have since been o bserved. We observed surprisingly good generalization performance on this matrix extrapolation task, both within domains (e.g., new users and new movies drawn f rom the same distribution used for training) and even across domains (e.g., pred icting music ratings after training on movie ratings).
****

Learning Memory Access Patterns
Milad Hashemi, Kevin Swersky, Jamie Smith, Grant Ayers, Heiner Litz, Jichuan Cha ng, Christos Kozyrakis, Parthasarathy Ranganathan
The explosion in workload complexity and the recent slow-down in Moore’s law sca ling call for new approaches towards efficient computing. Researchers are now be ginning to use recent advances in machine learning in software optimizations; au gmenting or replacing traditional heuristics and data structures. However, the s

pace of machine learning for computer hardware architecture is only lightly expl
ored. In this paper, we demonstrate the potential of deep learning to address th
e von Neumann bottleneck of memory performance. We focus on the critical problem
 of learning memory access patterns, with the goal of constructing accurate and
efficient memory prefetchers. We relate contemporary prefetching strategies to n
-gram models in natural language processing, and show how recurrent neural netwo
rks can serve as a drop-in replacement. On a suite of challenging benchmark data
sets, we find that neural networks consistently demonstrate superior performance
 in terms of precision and recall. This work represents the first step towards p
ractical neural-network based prefetching, and opens a wide range of exciting di
rections for machine learning in computer architecture research.
****

Fairness Without Demographics in Repeated Loss Minimization
Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, Percy Liang
Machine learning models (e.g., speech recognizers) trained on average loss suffe
r from representation disparity—minority groups (e.g., non-native speakers) carr
y less weight in the training objective, and thus tend to suffer higher loss. Wo
rse, as model accuracy affects user retention, a minority group can shrink over
time. In this paper, we first show that the status quo of empirical risk minimiz
ation (ERM) amplifies representation disparity over time, which can even turn in
itially fair models unfair. To mitigate this, we develop an approach based on di
stributionally robust optimization (DRO), which minimizes the worst case risk ov
er all distributions close to the empirical distribution. We prove that this app
roach controls the risk of the minority group at each time step, in the spirit o
f Rawlsian distributive justice, while remaining oblivious to the identity of th
e groups. We demonstrate that DRO prevents disparity amplification on examples w
here ERM fails, and show improvements in minority group user satisfaction in a r
eal-world text autocomplete task.
****

Multicalibration: Calibration for the (Computationally-Identifiable) Masses
Ursula Hebert-Johnson, Michael Kim, Omer Reingold, Guy Rothblum
We develop and study multicalibration as a new measure of fairness in machine le
arning that aims to mitigate inadvertent or malicious discrimination that is int
roduced at training time (even from ground truth data). Multicalibration guarant
ees meaningful (calibrated) predictions for every subpopulation that can be iden
tified within a specified class of computations. The specified class can be quit
e rich; in particular, it can contain many overlapping subgroups of a protected
group. We demonstrate that in many settings this strong notion of protection fro
m discrimination is provably attainable and aligned with the goal of obtaining a
ccurate predictions. Along the way, we present algorithms for learning a multica
librated predictor, study the computational complexity of this task, and illustr
ate tight connections to the agnostic learning model.
****

Recurrent Predictive State Policy Networks
Ahmed Hefny, Zita Marinho, Wen Sun, Siddhartha Srinivasa, Geoffrey Gordon
We introduce Recurrent Predictive State Policy(RPSP) networks, a recurrent archi
tecture that brings insights from predictive state representations to reinforcem
ent learning in partially ob-servable environments. Predictive state policy netw
orks consist of a recursive filter, which keeps track of a belief about the stat
e of the environment, and a reactive policy that directly maps beliefs to action
s, to maximize the cumulative reward. The recursive filter leverages predictive
state representations (PSRs) (Rosencrantz & Gordon, 2004; Sun et al., 2016) by m
odeling predictive state{—}a prediction of the distribution of future observatio
ns conditioned on history and future actions.This representation gives rise to a
 rich class of statistically consistent algorithms (Hefny et al.,2017) to initia
lize the recursive filter. Predictive stats serves as an equivalent representati
on of a belief state. Therefore, the policy component of the RPSP-network can be
 purely reactive, simplifying training while still allowing optimal behavior. Mo
reover, we use the PSR interpretation during training as well, by incorporating
prediction error in the loss function. The entire network (recursive filter and

reactive policy) is still differentiable and can be trained using gradient-based methods. We optimize our policy using a combination of policy gradient based on rewards (Williams, 1992)and gradient descent based on prediction error.We show the efficacy of RPSP-networks on a set of robotic control tasks from OpenAI Gym. We empirically show that RPSP-networks perform well compared with memory-preserving networks such as GRUs, as well as finite memory models, being the overall best performing method.
****

Learning unknown ODE models with Gaussian processes
Markus Heinonen, Cagatay Yildiz, Henrik Mannerström, Jukka Intosalmi, Harri Lähdesmäki
In conventional ODE modelling coefficients of an equation driving the system state forward in time are estimated. However, for many complex systems it is practically impossible to determine the equations or interactions governing the underlying dynamics. In these settings, parametric ODE model cannot be formulated. Here, we overcome this issue by introducing a novel paradigm of nonparametric ODE modelling that can learn the underlying dynamics of arbitrary continuous-time systems without prior knowledge. We propose to learn non-linear, unknown differential functions from state observations using Gaussian process vector fields within the exact ODE formalism. We demonstrate the model’s capabilities to infer dynamics from sparse data and to simulate the system forward into future.
****

Orthogonal Recurrent Neural Networks with Scaled Cayley Transform
Kyle Helfrich, Devin Willmott, Qiang Ye
Recurrent Neural Networks (RNNs) are designed to handle sequential data but suffer from vanishing or exploding gradients. Recent work on Unitary Recurrent Neural Networks (uRNNs) have been used to address this issue and in some cases, exceed the capabilities of Long Short-Term Memory networks (LSTMs). We propose a simpler and novel update scheme to maintain orthogonal recurrent weight matrices without using complex valued matrices. This is done by parametrizing with a skew-symmetric matrix using the Cayley transform; such a parametrization is unable to represent matrices with negative one eigenvalues, but this limitation is overcome by scaling the recurrent weight matrix by a diagonal matrix consisting of ones and negative ones. The proposed training scheme involves a straightforward gradient calculation and update step. In several experiments, the proposed scaled Cayley orthogonal recurrent neural network (scoRNN) achieves superior results with fewer trainable parameters than other unitary RNNs.
****

Fast Bellman Updates for Robust MDPs
Chin Pang Ho, Marek Petrik, Wolfram Wiesemann
We describe two efficient, and exact, algorithms for computing Bellman updates in robust Markov decision processes (MDPs). The first algorithm uses a homotopy continuation method to compute updates for L1-constrained s,a-rectangular ambiguity sets. It runs in quasi-linear time for plain L1-norms and also generalizes to weighted L1-norms. The second algorithm uses bisection to compute updates for robust MDPs with s-rectangular ambiguity sets. This algorithm, when combined with the homotopy method, also has a quasi-linear runtime. Unlike previous methods, our algorithms compute the primal solution in addition to the optimal objective value, which makes them useful in policy iteration methods. Our experimental results indicate that the proposed methods are over 1,000 times faster than Gurobi, a state-of-the-art commercial optimization package, for small instances, and the performance gap grows considerably with problem size.
****

CyCADA: Cycle-Consistent Adversarial Domain Adaptation
Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, Trevor Darrell
Domain adaptation is critical for success in new, unseen environments. Adversarial adaptation models have shown tremendous progress towards adapting to new environments by focusing either on discovering domain invariant representations or by mapping between unpaired image domains. While feature space methods are diffic

ult to interpret and sometimes fail to capture pixel-level and low-level domain shifts, image space methods sometimes fail to incorporate high level semantic kn owledge relevant for the end task. We propose a model which adapts between domai ns using both generative image space alignment and latent representation space a lignment. Our approach, Cycle-Consistent Adversarial Domain Adaptation (CyCADA), guides transfer between domains according to a specific discriminatively traine d task and avoids divergence by enforcing consistency of the relevant semantics before and after adaptation. We evaluate our method on a variety of visual recog nition and prediction settings, including digit classification and semantic segm entation of road scenes, advancing state-of-the-art performance for unsupervised adaptation from synthetic to real world driving domains.
****
Sound Abstraction and Decomposition of Probabilistic Programs
Steven Holtzen, Guy Broeck, Todd Millstein
Probabilistic programming languages are a flexible tool for specifying statistic al models, but this flexibility comes at the cost of efficient analysis. It is c urrently difficult to compactly represent the subtle independence properties of a probabilistic program, and exploit independence properties to decompose infere nce. Classical graphical model abstractions do capture some properties of the un derlying distribution, enabling inference algorithms to operate at the level of the graph topology. However, we observe that graph-based abstractions are often too coarse to capture interesting properties of programs. We propose a form of s ound abstraction for probabilistic programs wherein the abstractions are themsel ves simplified programs. We provide a theoretical foundation for these abstracti ons, as well as an algorithm to generate them. Experimentally, we also illustrat e the practical benefits of our framework as a tool to decompose probabilistic p rogram inference.
****
Gradient Primal-Dual Algorithm Converges to Second-Order Stationary Solution for Nonconvex Distributed Optimization Over Networks
Mingyi Hong, Meisam Razaviyayn, Jason Lee
In this work, we study two first-order primal-dual based algorithms, the Gradien t Primal-Dual Algorithm (GPDA) and the Gradient Alternating Direction Method of Multipliers (GADMM), for solving a class of linearly constrained non-convex opti mization problems. We show that with random initialization of the primal and dua l variables, both algorithms are able to compute second-order stationary solutio ns (ss2) with probability one. This is the first result showing that primal-dual algorithm is capable of finding ss2 when only using first-order information; it also extends the existing results for first-order, but {primal-only} algorithms . An important implication of our result is that it also gives rise to the first global convergence result to the ss2, for two classes of unconstrained distribu ted non-convex learning problems over multi-agent networks.
****
Variational Bayesian dropout: pitfalls and fixes
Jiri Hron, Alex Matthews, Zoubin Ghahramani
Dropout, a stochastic regularisation technique for training of neural networks, has recently been reinterpreted as a specific type of approximate inference algo rithm for Bayesian neural networks. The main contribution of the reinterpretatio n is in providing a theoretical framework useful for analysing and extending the algorithm. We show that the proposed framework suffers from several issues; fro m undefined or pathological behaviour of the true posterior related to use of im proper priors, to an ill-defined variational objective due to singularity of the approximating distribution relative to the true posterior. Our analysis of the improper log uniform prior used in variational Gaussian dropout suggests the pat hologies are generally irredeemable, and that the algorithm still works only bec ause the variational formulation annuls some of the pathologies. To address the singularity issue, we proffer Quasi-KL (QKL) divergence, a new approximate infer ence objective for approximation of high-dimensional distributions. We show that motivations for variational Bernoulli dropout based on discretisation and noise have QKL as a limit. Properties of QKL are studied both theoretically and on a

simple practical example which shows that the QKL-optimal approximation of a ful l rank Gaussian with a degenerate one naturally leads to the Principal Component Analysis solution.
****

Does Distributionally Robust Supervised Learning Give Robust Classifiers?
Weihua Hu, Gang Niu, Issei Sato, Masashi Sugiyama
Distributionally Robust Supervised Learning (DRSL) is necessary for building rel iable machine learning systems. When machine learning is deployed in the real wo rld, its performance can be significantly degraded because test data may follow a different distribution from training data. DRSL with f-divergences explicitly considers the worst-case distribution shift by minimizing the adversarially rewe ighted training loss. In this paper, we analyze this DRSL, focusing on the class ification scenario. Since the DRSL is explicitly formulated for a distribution s hift scenario, we naturally expect it to give a robust classifier that can aggre ssively handle shifted distributions. However, surprisingly, we prove that the D RSL just ends up giving a classifier that exactly fits the given training distri bution, which is too pessimistic. This pessimism comes from two sources: the par ticular losses used in classification and the fact that the variety of distribut ions to which the DRSL tries to be robust is too wide. Motivated by our analysis , we propose simple DRSL that overcomes this pessimism and empirically demonstra te its effectiveness.
****

Dissipativity Theory for Accelerating Stochastic Variance Reduction: A Unified A nalysis of SVRG and Katyusha Using Semidefinite Programs
Bin Hu, Stephen Wright, Laurent Lessard
Techniques for reducing the variance of gradient estimates used in stochastic pr ogramming algorithms for convex finite-sum problems have received a great deal o f attention in recent years. By leveraging dissipativity theory from control, we provide a new perspective on two important variance-reduction algorithms: SVRG and its direct accelerated variant Katyusha. Our perspective provides a physical ly intuitive understanding of the behavior of SVRG-like methods via a principle of energy conservation. The tools discussed here allow us to automate the conver gence analysis of SVRG-like methods by capturing their essential properties in s mall semidefinite programs amenable to standard analysis and computational techn iques. Our approach recovers existing convergence results for SVRG and Katyusha and generalizes the theory to alternative parameter choices. We also discuss how our approach complements the linear coupling technique. Our combination of pers pectives leads to a better understanding of accelerated variance-reduced stochas tic methods for finite-sum problems.
****

Near Optimal Frequent Directions for Sketching Dense and Sparse Matrices
Zengfeng Huang
Given a large matrix $A\in\real^{n\times d}$, we consider the problem of computi ng a sketch matrix $B\in\real^{\ell\times d}$ which is significantly smaller tha n but still well approximates A. We are interested in minimizing the covarianc e error $\norm{A^TA-B^TB}_2.$We consider the problems in the streaming model, wh ere the algorithm can only make one pass over the input with limited working spa ce. The popular Frequent Directions algorithm of Liberty (2013) and its variants achieve optimal space-error tradeoff. However, whether the running time can be improved remains an unanswered question.In this paper, we almost settle the time complexity of this problem. In particular, we provide new space-optimal algorit hms with faster running times. Moreover, we also show that the running times of our algorithms are near-optimal unless the state-of-the-art running time of matr ix multiplication can be improved significantly.
****

Learning Deep ResNet Blocks Sequentially using Boosting Theory
Furong Huang, Jordan Ash, John Langford, Robert Schapire
We prove a multi-channel telescoping sum boosting theory for the ResNet architec tures which simultaneously creates a new technique for boosting over features (i n contrast with labels) and provides a new algorithm for ResNet-style architectu

res. Our proposed training algorithm, BoostResNet, is particularly suitable in n
on-differentiable architectures. Our method only requires the relatively inexpen
sive sequential training of T “shallow ResNets”. We prove that the training er
ror decays exponentially with the depth T if the weak module classifiers that
we train perform slightly better than some weak baseline. In other words, we pro
pose a weak learning condition and prove a boosting theory for ResNet under the
weak learning condition. A generalization error bound based on margin theory is
proved and suggests that ResNet could be resistant to overfitting using a networ
k with l_1 norm bounded weights.
****

Learning Hidden Markov Models from Pairwise Co-occurrences with Application to T
opic Modeling
Kejun Huang, Xiao Fu, Nicholas Sidiropoulos
We present a new algorithm for identifying the transition and emission probabili
ties of a hidden Markov model (HMM) from the emitted data. Expectation-maximizat
ion becomes computationally prohibitive for long observation records, which are
often required for identification. The new algorithm is particularly suitable fo
r cases where the available sample size is large enough to accurately estimate s
econd-order output probabilities, but not higher-order ones. We show that if one
 is only able to obtain a reliable estimate of the pairwise co-occurrence probab
ilities of the emissions, it is still possible to uniquely identify the HMM if t
he emission probability is sufficiently scattered. We apply our method to hidden
 topic Markov modeling, and demonstrate that we can learn topics with higher qua
lity if documents are modeled as observations of HMMs sharing the same emission
(topic) probability, compared to the simple but widely used bag-of-words model.
****

Neural Autoregressive Flows
Chin-Wei Huang, David Krueger, Alexandre Lacoste, Aaron Courville
Normalizing flows and autoregressive models have been successfully combined to p
roduce state-of-the-art results in density estimation, via Masked Autoregressive
 Flows (MAF) (Papamakarios et al., 2017), and to accelerate state-of-the-art Wav
eNet-based speech synthesis to 20x faster than real-time (Oord et al., 2017), vi
a Inverse Autoregressive Flows (IAF) (Kingma et al., 2016). We unify and general
ize these approaches, replacing the (conditionally) affine univariate transforma
tions of MAF/IAF with a more general class of invertible univariate transformati
ons expressed as monotonic neural networks. We demonstrate that the proposed neu
ral autoregressive flows (NAF) are universal approximators for continuous probab
ility distributions, and their greater expressivity allows them to better captur
e multimodal target distributions. Experimentally, NAF yields state-of-the-art p
erformance on a suite of density estimation tasks and outperforms IAF in variati
onal autoencoders trained on binarized MNIST.
****

Topological mixture estimation
Steve Huntsman
We introduce topological mixture estimation, a completely nonparametric and comp
utationally efficient solution to the problem of estimating a one-dimensional mi
xture with generic unimodal components. We repeatedly perturb the unimodal decom
position of Baryshnikov and Ghrist to produce a topologically and information-th
eoretically optimal unimodal mixture. We also detail a smoothing process that op
timally exploits topological persistence of the unimodal category in a natural w
ay when working directly with sample data. Finally, we illustrate these techniqu
es through examples.
****

Decoupled Parallel Backpropagation with Convergence Guarantee
Zhouyuan Huo, Bin Gu, Yang, Heng Huang
Backpropagation algorithm is indispensable for the training of feedforward neura
l networks. It requires propagating error gradients sequentially from the output
 layer all the way back to the input layer. The backward locking in backpropagat
ion algorithm constrains us from updating network layers in parallel and fully l
everaging the computing resources. Recently, several algorithms have been propos

ed for breaking the backward locking. However, their performances degrade seriou sly when networks are deep. In this paper, we propose decoupled parallel backpro pagation algorithm for deep learning optimization with convergence guarantee. Fi rstly, we decouple the backpropagation algorithm using delayed gradients, and sh ow that the backward locking is removed when we split the networks into multiple modules. Then, we utilize decoupled parallel backpropagation in two stochastic methods and prove that our method guarantees convergence to critical points for the non-convex problem. Finally, we perform experiments for training deep convol utional neural networks on benchmark datasets. The experimental results not only confirm our theoretical analysis, but also demonstrate that the proposed method can achieve significant speedup without loss of accuracy.

****

Using Reward Machines for High-Level Task Specification and Decomposition in Rei nforcement Learning

Rodrigo Toro Icarte, Toryn Klassen, Richard Valenzano, Sheila McIlraith

In this paper we propose Reward Machines {—} a type of finite state machine that supports the specification of reward functions while exposing reward function s tructure to the learner and supporting decomposition. We then present Q-Learning for Reward Machines (QRM), an algorithm which appropriately decomposes the rewa rd machine and uses off-policy q-learning to simultaneously learn subpolicies fo r the different components. QRM is guaranteed to converge to an optimal policy i n the tabular case, in contrast to Hierarchical Reinforcement Learning methods w hich might converge to suboptimal policies. We demonstrate this behavior experim entally in two discrete domains. We also show how function approximation methods like neural networks can be incorporated into QRM, and that doing so can find b etter policies more quickly than hierarchical methods in a domain with a continu ous state space.

****

Deep Variational Reinforcement Learning for POMDPs

Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, Shimon Whiteson

Many real-world sequential decision making problems are partially observable by nature, and the environment model is typically unknown. Consequently, there is g reat need for reinforcement learning methods that can tackle such problems given only a stream of rewards and incomplete and noisy observations. In this paper, we propose deep variational reinforcement learning (DVRL), which introduces an i nductive bias that allows an agent to learn a generative model of the environmen t and perform inference in that model to effectively aggregate the available inf ormation. We develop an n-step approximation to the evidence lower bound (ELBO), allowing the model to be trained jointly with the policy. This ensures that the latent state representation is suitable for the control task. In experiments on Mountain Hike and flickering Atari we show that our method outperforms previous approaches relying on recurrent neural networks to encode the past.

****

Attention-based Deep Multiple Instance Learning

Maximilian Ilse, Jakub Tomczak, Max Welling

Multiple instance learning (MIL) is a variation of supervised learning where a s ingle class label is assigned to a bag of instances. In this paper, we state the MIL problem as learning the Bernoulli distribution of the bag label where the b ag label probability is fully parameterized by neural networks. Furthermore, we propose a neural network-based permutation-invariant aggregation operator that c orresponds to the attention mechanism. Notably, an application of the proposed a ttention-based operator provides insight into the contribution of each instance to the bag label. We show empirically that our approach achieves comparable perf ormance to the best MIL methods on benchmark MIL datasets and it outperforms oth er methods on a MNIST-based MIL dataset and two real-life histopathology dataset s without sacrificing interpretability.

****

Black-box Adversarial Attacks with Limited Queries and Information

Andrew Ilyas, Logan Engstrom, Anish Athalye, Jessy Lin

Current neural network-based classifiers are susceptible to adversarial examples

even in the black-box setting, where the attacker only has query access to the model. In practice, the threat model for real-world systems is often more restrictive than the typical black-box model where the adversary can observe the full output of the network on arbitrarily many chosen inputs. We define three realistic threat models that more accurately characterize many real-world classifiers: the query-limited setting, the partial-information setting, and the label-only setting. We develop new attacks that fool classifiers under these more restrictive threat models, where previous methods would be impractical or ineffective. We demonstrate that our methods are effective against an ImageNet classifier under our proposed threat models. We also demonstrate a targeted black-box attack against a commercial classifier, overcoming the challenges of limited query access, partial information, and other practical issues to break the Google Cloud Vision API.

****

Analysis of Minimax Error Rate for Crowdsourcing and Its Application to Worker Clustering Model

Hideaki Imamura, Issei Sato, Masashi Sugiyama

While crowdsourcing has become an important means to label data, there is great interest in estimating the ground truth from unreliable labels produced by crowd workers. The Dawid and Skene (DS) model is one of the most well-known models in the study of crowdsourcing. Despite its practical popularity, theoretical error analysis for the DS model has been conducted only under restrictive assumptions on class priors, confusion matrices, or the number of labels each worker provides. In this paper, we derive a minimax error rate under more practical setting for a broader class of crowdsourcing models including the DS model as a special case. We further propose the worker clustering model, which is more practical than the DS model under real crowdsourcing settings. The wide applicability of our theoretical analysis allows us to immediately investigate the behavior of this proposed model, which can not be analyzed by existing studies. Experimental results showed that there is a strong similarity between the lower bound of the minimax error rate derived by our theoretical analysis and the empirical error of the estimated value.

****

Improving Regression Performance with Distributional Losses

Ehsan Imani, Martha White

There is growing evidence that converting targets to soft targets in supervised learning can provide considerable gains in performance. Much of this work has considered classification, converting hard zero-one values to soft labels—such as by adding label noise, incorporating label ambiguity or using distillation. In parallel, there is some evidence from a regression setting in reinforcement learning that learning distributions can improve performance. In this work, we investigate the reasons for this improvement, in a regression setting. We introduce a novel distributional regression loss, and similarly find it significantly improves prediction accuracy. We investigate several common hypotheses, around reducing overfitting and improved representations. We instead find evidence for an alternative hypothesis: this loss is easier to optimize, with better behaved gradients, resulting in improved generalization. We provide theoretical support for this alternative hypothesis, by characterizing the norm of the gradients of this loss.

****

Deep Density Destructors

David Inouye, Pradeep Ravikumar

We propose a unified framework for deep density models by formally defining density destructors. A density destructor is an invertible function that transforms a given density to the uniform density—essentially destroying any structure in the original density. This destructive transformation generalizes Gaussianization via ICA and more recent autoregressive models such as MAF and Real NVP. Informally, this transformation can be seen as a generalized whitening procedure or a multivariate generalization of the univariate CDF function. Unlike Gaussianization, our destructive transformation has the elegant property that the density func

tion is equal to the absolute value of the Jacobian determinant. Thus, each laye r of a deep density can be seen as a shallow density—uncovering a fundamental co nnection between shallow and deep densities. In addition, our framework provides a common interface for all previous methods enabling them to be systematically combined, evaluated and improved. Leveraging the connection to shallow densities , we also propose a novel tree destructor based on tree densities and an image-s pecific destructor based on pixel locality. We illustrate our framework on a 2D dataset, MNIST, and CIFAR-10. Code is available on first author's website.
****

Unbiased Objective Estimation in Predictive Optimization
Shinji Ito, Akihiro Yabe, Ryohei Fujimaki
For data-driven decision-making, one promising approach, called predictive optim ization, is to solve maximization problems i n which the objective function to b e maximized is estimated from data. Predictive optimization, however, suffers fr om the problem of a calculated optimal solution's being evaluated too optimistic ally, i.e., the value of the objective function is overestimated. This paper inv estigates such optimistic bias and presents two methods for correcting it. The f irst, which is analogous to cross-validation, successfully corrects the optimist ic bias but results in underestimation of the true value. Our second method empl oys resampling techniques to avoid both overestimation and underestimation. We s how that the second method, referred to as the parameter perturbation method, ac hieves asymptotically unbiased estimation. Empirical results for both artificial and real-world datasets demonstrate that our proposed approach successfully cor rects the optimistic bias.
****

Anonymous Walk Embeddings
Sergey Ivanov, Evgeny Burnaev
The task of representing entire graphs has seen a surge of prominent results, ma inly due to learning convolutional neural networks (CNNs) on graph-structured da ta. While CNNs demonstrate state-of-the-art performance in graph classification task, such methods are supervised and therefore steer away from the original pro blem of network representation in task-agnostic manner. Here, we coherently prop ose an approach for embedding entire graphs and show that our feature representa tions with SVM classifier increase classification accuracy of CNN algorithms and traditional graph kernels. For this we describe a recently discovered graph obj ect, anonymous walk, on which we design task-independent algorithms for learning graph representations in explicit and distributed way. Overall, our work repres ents a new scalable unsupervised learning of state-of-the-art representations of entire graphs.
****

Learning Binary Latent Variable Models: A Tensor Eigenpair Approach
Ariel Jaffe, Roi Weiss, Boaz Nadler, Shai Carmi, Yuval Kluger
Latent variable models with hidden binary units appear in various applications. Learning such models, in particular in the presence of noise, is a challenging c omputational problem. In this paper we propose a novel spectral approach to this problem, based on the eigenvectors of both the second order moment matrix and t hird order moment tensor of the observed data. We prove that under mild non-dege neracy conditions, our method consistently estimates the model parameters at the optimal parametric rate. Our tensor-based method generalizes previous orthogona l tensor decomposition approaches, where the hidden units were assumed to be eit her statistically independent or mutually exclusive. We illustrate the consisten cy of our method on simulated data and demonstrate its usefulness in learning a common model for population mixtures in genetics.
****

Firing Bandits: Optimizing Crowdfunding
Lalit Jain, Kevin Jamieson
In this paper, we model the problem of optimizing crowdfunding platforms, such a s the non-profit Kiva or for-profit KickStarter, as a variant of the multi-armed bandit problem. In our setting, Bernoulli arms emit no rewards until their cumu lative number of successes over any number of trials exceeds a fixed threshold a

nd then provides no additional reward for any additional trials - a process remi
niscent to that of a neuron firing once it reaches the action potential and then
 saturates. In the spirit of an infinite armed bandit problem, the player can ad
d new arms whose expected probability of success is drawn iid from an unknown di
stribution – this endless supply of projects models the harsh reality that the n
umber of projects seeking funding greatly exceeds the total capital available by
 lenders. Crowdfunding platforms naturally fall under this setting where the arm
s are potential projects, and their probability of success is the probability th
at a potential funder decides to fund it after reviewing it. The goal is to play
 arms (prioritize the display of projects on a webpage) to maximize the number o
f arms that reach the firing threshold (meet their goal amount) using as few tot
al trials (number of impressions) as possible over all the played arms. We provi
de an algorithm for this setting and prove sublinear regret bounds.
****

Differentially Private Matrix Completion Revisited
Prateek Jain, Om Dipakbhai Thakkar, Abhradeep Thakurta
We provide the first provably joint differentially private algorithm with formal
 utility guarantees for the problem of user-level privacy-preserving collaborati
ve filtering. Our algorithm is based on the Frank-Wolfe method, and it consisten
tly estimates the underlying preference matrix as long as the number of users $m
$ is $\omega(n^{5/4})$, where n is the number of items, and each user provides
 her preference for at least $\sqrt{n}$ randomly selected items. Along the way,
we provide an optimal differentially private algorithm for singular vector compu
tation, based on the celebrated Oja’s method, that provides significant savings
in terms of space and time while operating on sparse matrices. We also empirical
ly evaluate our algorithm on a suite of datasets, and show that it consistently
outperforms the state-of-the-art private algorithms.
****

Video Prediction with Appearance and Motion Conditions
Yunseok Jang, Gunhee Kim, Yale Song
Video prediction aims to generate realistic future frames by learning dynamic vi
sual patterns. One fundamental challenge is to deal with future uncertainty: How
 should a model behave when there are multiple correct, equally probable future?
 We propose an Appearance-Motion Conditional GAN to address this challenge. We p
rovide appearance and motion information as conditions that specify how the futu
re may look like, reducing the level of uncertainty. Our model consists of a gen
erator, two discriminators taking charge of appearance and motion pathways, and
a perceptual ranking module that encourages videos of similar conditions to look
 similar. To train our model, we develop a novel conditioning scheme that consis
ts of different combinations of appearance and motion conditions. We evaluate ou
r model using facial expression and human action datasets and report favorable r
esults compared to existing methods.
****

Pathwise Derivatives Beyond the Reparameterization Trick
Martin Jankowiak, Fritz Obermeyer
We observe that gradients computed via the reparameterization trick are in direc
t correspondence with solutions of the transport equation in the formalism of op
timal transport. We use this perspective to compute (approximate) pathwise gradi
ents for probability distributions not directly amenable to the reparameterizati
on trick: Gamma, Beta, and Dirichlet. We further observe that when the reparamet
erization trick is applied to the Cholesky-factorized multivariate Normal distri
bution, the resulting gradients are suboptimal in the sense of optimal transport
. We derive the optimal gradients and show that they have reduced variance in a
Gaussian Process regression task. We demonstrate with a variety of synthetic exp
eriments and stochastic variational inference tasks that our pathwise gradients
are competitive with other methods.
****

Detecting non-causal artifacts in multivariate linear regression models
Dominik Janzing, Bernhard Schölkopf
We consider linear models where d potential causes X_1,...,X_d are correlated wi

th one target quantity Y and propose a method to infer whether the association i
s causal or whether it is an artifact caused by overfitting or hidden common cau
ses. We employ the idea that in the former case the vector of regression coeffic
ients has ‘generic’ orientation relative to the covariance matrix Sigma_{XX} of
X. Using an ICA based model for confounding, we show that both confounding and o
verfitting yield regression vectors that concentrate mainly in the space of low
eigenvalues of Sigma_{XX}.
****

A Unified Framework for Structured Low-rank Matrix Learning
Pratik Jawanpuria, Bamdev Mishra
We consider the problem of learning a low-rank matrix, constrained to lie in a l
inear subspace, and introduce a novel factorization for modeling such matrices.
A salient feature of the proposed factorization scheme is it decouples the low-r
ank and the structural constraints onto separate factors. We formulate the optim
ization problem on the Riemannian spectrahedron manifold, where the Riemannian f
ramework allows to develop computationally efficient conjugate gradient and trus
t-region algorithms. Experiments on problems such as standard/robust/non-negativ
e matrix completion, Hankel matrix learning and multi-task learning demonstrate
the efficacy of our approach.
****

Efficient end-to-end learning for quantizable representations
Yeonwoo Jeong, Hyun Oh Song
Embedding representation learning via neural networks is at the core foundation
of modern similarity based search. While much effort has been put in developing
algorithms for learning binary hamming code representations for search efficienc
y, this still requires a linear scan of the entire dataset per each query and tr
ades off the search accuracy through binarization. To this end, we consider the
problem of directly learning a quantizable embedding representation and the spar
se binary hash code end-to-end which can be used to construct an efficient hash
table not only providing significant search reduction in the number of data but
also achieving the state of the art search accuracy outperforming previous state
of the art deep metric learning methods. We also show that finding the optimal
sparse binary hash code in a mini-batch can be computed exactly in polynomial ti
me by solving a minimum cost flow problem. Our results on Cifar-100 and on Image
Net datasets show the state of the art search accuracy in precision@k and NMI me
trics while providing up to 98X and 478X search speedup respectively over exhaus
tive linear search. The source code is available at https://github.com/maestroje
ong/Deep-Hash-Table-ICML18.
****

Exploring Hidden Dimensions in Accelerating Convolutional Neural Networks
Zhihao Jia, Sina Lin, Charles R. Qi, Alex Aiken
The past few years have witnessed growth in the computational requirements for t
raining deep convolutional neural networks. Current approaches parallelize train
ing onto multiple devices by applying a single parallelization strategy (e.g., d
ata or model parallelism) to all layers in a network. Although easy to reason ab
out, these approaches result in suboptimal runtime performance in large-scale di
stributed training, since different layers in a network may prefer different par
allelization strategies. In this paper, we propose layer-wise parallelism that a
llows each layer in a network to use an individual parallelization strategy. We
jointly optimize how each layer is parallelized by solving a graph search proble
m. Our evaluation shows that layer-wise parallelism outperforms state-of-the-art
approaches by increasing training throughput, reducing communication costs, ach
ieving better scalability to multiple GPUs, while maintaining original network a
ccuracy.
****

Feedback-Based Tree Search for Reinforcement Learning
Daniel Jiang, Emmanuel Ekwedike, Han Liu
Inspired by recent successes of Monte-Carlo tree search (MCTS) in a number of ar
tificial intelligence (AI) application domains, we propose a reinforcement learn
ing (RL) technique that iteratively applies MCTS on batches of small, finite-hor

izon versions of the original infinite-horizon Markov decision process. The terminal condition of the finite-horizon problems, or the leaf-node evaluator of the decision tree generated by MCTS, is specified using a combination of an estimated value function and an estimated policy function. The recommendations generated by the MCTS procedure are then provided as feedback in order to refine, through classification and regression, the leaf-node evaluator for the next iteration. We provide the first sample complexity bounds for a tree search-based RL algorithm. In addition, we show that a deep neural network implementation of the technique can create a competitive AI agent for the popular multi-player online battle arena (MOBA) game King of Glory.
****

Quickshift++: Provably Good Initializations for Sample-Based Mean Shift
Heinrich Jiang, Jennifer Jang, Samory Kpotufe
We provide initial seedings to the Quick Shift clustering algorithm, which approximate the locally high-density regions of the data. Such seedings act as more stable and expressive cluster-cores than the singleton modes found by Quick Shift. We establish statistical consistency guarantees for this modification. We then show strong clustering performance on real datasets as well as promising applications to image segmentation.
****

MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels
Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, Li Fei-Fei
Recent deep networks are capable of memorizing the entire data even when the labels are completely random. To overcome the overfitting on corrupted labels, we propose a novel technique of learning another neural network, called MentorNet, to supervise the training of the base deep networks, namely, StudentNet. During training, MentorNet provides a curriculum (sample weighting scheme) for StudentNet to focus on the sample the label of which is probably correct. Unlike the existing curriculum that is usually predefined by human experts, MentorNet learns a data-driven curriculum dynamically with StudentNet. Experimental results demonstrate that our approach can significantly improve the generalization performance of deep networks trained on corrupted training data. Notably, to the best of our knowledge, we achieve the best-published result on WebVision, a large benchmark containing 2.2 million images of real-world noisy labels.
****

The Weighted Kendall and High-order Kernels for Permutations
Yunlong Jiao, Jean-Philippe Vert
We propose new positive definite kernels for permutations. First we introduce a weighted version of the Kendall kernel, which allows to weight unequally the contributions of different item pairs in the permutations depending on their ranks. Like the Kendall kernel, we show that the weighted version is invariant to relabeling of items and can be computed efficiently in O(n ln(n)) operations, where n is the number of items in the permutation. Second, we propose a supervised approach to learn the weights by jointly optimizing them with the function estimated by a kernel machine. Third, while the Kendall kernel considers pairwise comparison between items, we extend it by considering higher-order comparisons among tuples of items and show that the supervised approach of learning the weights can be systematically generalized to higher-order permutation kernels.
****

Junction Tree Variational Autoencoder for Molecular Graph Generation
Wengong Jin, Regina Barzilay, Tommi Jaakkola
We seek to automate the design of molecules based on specific chemical properties. In computational terms, this task involves continuous embedding and generation of molecular graphs. Our primary contribution is the direct realization of molecular graphs, a task previously approached by generating linear SMILES strings instead of graphs. Our junction tree variational autoencoder generates molecular graphs in two phases, by first generating a tree-structured scaffold over chemical substructures, and then combining them into a molecule with a graph message passing network. This approach allows us to incrementally expand molecules while

maintaining chemical validity at every step. We evaluate our model on multiple tasks ranging from molecular generation to optimization. Across these tasks, our model outperforms previous state-of-the-art baselines by a significant margin.
****

Network Global Testing by Counting Graphlets
Jiashun Jin, Zheng Ke, Shengming Luo
Consider a large social network with possibly severe degree heterogeneity and mixed-memberships. We are interested in testing whether the network has only one community or there are more than one communities. The problem is known to be non-trivial, partially due to the presence of severe degree heterogeneity. We construct a class of test statistics using the numbers of short paths and short cycles, and the key to our approach is a general framework for canceling the effects of degree heterogeneity. The tests compare favorably with existing methods. We support our methods with careful analysis and numerical study with simulated data and a real data example.
****

Regret Minimization for Partially Observable Deep Reinforcement Learning
Peter Jin, Kurt Keutzer, Sergey Levine
Deep reinforcement learning algorithms that estimate state and state-action value functions have been shown to be effective in a variety of challenging domains, including learning control strategies from raw image pixels. However, algorithms that estimate state and state-action value functions typically assume a fully observed state and must compensate for partial observations by using finite length observation histories or recurrent networks. In this work, we propose a new deep reinforcement learning algorithm based on counterfactual regret minimization that iteratively updates an approximation to an advantage-like function and is robust to partially observed state. We demonstrate that this new algorithm can substantially outperform strong baseline methods on several partially observed reinforcement learning tasks: learning first-person 3D navigation in Doom and Minecraft, and acting in the presence of partially observed objects in Doom and Pong.
****

WSNet: Compact and Efficient Networks Through Weight Sampling
Xiaojie Jin, Yingzhen Yang, Ning Xu, Jianchao Yang, Nebojsa Jojic, Jiashi Feng, Shuicheng Yan
We present a new approach and a novel architecture, termed WSNet, for learning compact and efficient deep neural networks. Existing approaches conventionally learn full model parameters independently and then compress them via ad hoc processing such as model pruning or filter factorization. Alternatively, WSNet proposes learning model parameters by sampling from a compact set of learnable parameters, which naturally enforces parameter sharing throughout the learning process. We demonstrate that such a novel weight sampling approach (and induced WSNet) promotes both weights and computation sharing favorably. By employing this method, we can more efficiently learn much smaller networks with competitive performance compared to baseline networks with equal numbers of convolution filters. Specifically, we consider learning compact and efficient 1D convolutional neural networks for audio classification. Extensive experiments on multiple audio classification datasets verify the effectiveness of WSNet. Combined with weight quantization, the resulted models are up to 180x smaller and theoretically up to 16x faster than the well-established baselines, without noticeable performance drop.
****

Large-Scale Cox Process Inference using Variational Fourier Features
ST John, James Hensman
Gaussian process modulated Poisson processes provide a flexible framework for modeling spatiotemporal point patterns. So far this had been restricted to one dimension, binning to a pre-determined grid, or small data sets of up to a few thousand data points. Here we introduce Cox process inference based on Fourier features. This sparse representation induces global rather than local constraints on the function space and is computationally efficient. This allows us to formulate a grid-free approximation that scales well with the number of data points and t

he size of the domain. We demonstrate that this allows MCMC approximations to th e non-Gaussian posterior. In practice, we find that Fourier features have more c onsistent optimization behavior than previous approaches. Our approximate Bayesi an method can fit over 100 000 events with complex spatiotemporal patterns in th ree dimensions on a single GPU.
****

Composite Functional Gradient Learning of Generative Adversarial Models
Rie Johnson, Tong Zhang
This paper first presents a theory for generative adversarial methods that does not rely on the traditional minimax formulation. It shows that with a strong dis criminator, a good generator can be learned so that the KL divergence between th e distributions of real data and generated data improves after each functional g radient step until it converges to zero. Based on the theory, we propose a new s table generative adversarial method. A theoretical insight into the original GAN from this new viewpoint is also provided. The experiments on image generation s how the effectiveness of our new method.
****

Kronecker Recurrent Units
Cijo Jose, Moustapha Cisse, Francois Fleuret
Our work addresses two important issues with recurrent neural networks: (1) they are over-parametrized, and (2) the recurrent weight matrix is ill-conditioned. The former increases the sample complexity of learning and the training time. Th e latter causes the vanishing and exploding gradient problem. We present a flexi ble recurrent neural network model called Kronecker Recurrent Units (KRU). KRU a chieves parameter efficiency in RNNs through a Kronecker factored recurrent matr ix. It overcomes the ill-conditioning of the recurrent matrix by enforcing soft unitary constraints on the factors. Thanks to the small dimensionality of the fa ctors, maintaining these constraints is computationally efficient. Our experimen tal results on seven standard data-sets reveal that KRU can reduce the number of parameters by three orders of magnitude in the recurrent weight matrix compared to the existing recurrent models, without trading the statistical performance. These results in particular show that while there are advantages in having a hig h dimensional recurrent space, the capacity of the recurrent part of the model c an be dramatically reduced.
****

Fast Decoding in Sequence Models Using Discrete Latent Variables
Lukasz Kaiser, Samy Bengio, Aurko Roy, Ashish Vaswani, Niki Parmar, Jakob Uszkor eit, Noam Shazeer
Autoregressive sequence models based on deep neural networks, such as RNNs, Wave net and Transformer are the state-of-the-art on many tasks. However, they lack p arallelism and are thus slow for long sequences. RNNs lack parallelism both duri ng training and decoding, while architectures like WaveNet and Transformer are m uch more parallel during training, but still lack parallelism during decoding. W e present a method to extend sequence models using discrete latent variables tha t makes decoding much more parallel. The main idea behind this approach is to fi rst autoencode the target sequence into a shorter discrete latent sequence, whic h is generated autoregressively, and finally decode the full sequence from this shorter latent sequence in a parallel manner. To this end, we introduce a new me thod for constructing discrete latent variables and compare it with previously i ntroduced methods. Finally, we verify that our model works on the task of neural machine translation, where our models are an order of magnitude faster than com parable autoregressive models and, while lower in BLEU than purely autoregressiv e models, better than previously proposed non-autogregressive translation.
****

Kernel Recursive ABC: Point Estimation with Intractable Likelihood
Takafumi Kajihara, Motonobu Kanagawa, Keisuke Yamazaki, Kenji Fukumizu
We propose a novel approach to parameter estimation for simulator-based statisti cal models with intractable likelihood. Our proposed method involves recursive a pplication of kernel ABC and kernel herding to the same observed data. We provid e a theoretical explanation regarding why the approach works, showing (for the p

opulation setting) that, under a certain assumption, point estimates obtained wi th this method converge to the true parameter, as recursion proceeds. We have co nducted a variety of numerical experiments, including parameter estimation for a real-world pedestrian flow simulator, and show that in most cases our method ou tperforms existing approaches.

****

Efficient Neural Audio Synthesis

Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edw ard Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, Koray Kavukcuoglu

Sequential models achieve state-of-the-art results in audio, visual and textual domains with respect to both estimating the data distribution and generating des ired samples. Efficient sampling for this class of models at the cost of little to no loss in quality has however remained an elusive problem. With a focus on t ext-to-speech synthesis, we describe a set of general techniques for reducing sa mpling time while maintaining high output quality. We first describe a single-la yer recurrent neural network, the WaveRNN, with a dual softmax layer that matche s the quality of the state-of-the-art WaveNet model. The compact form of the net work makes it possible to generate 24 kHz 16-bit audio 4 times faster than real time on a GPU. Secondly, we apply a weight pruning technique to reduce the numbe r of weights in the WaveRNN. We find that, for a constant number of parameters, large sparse networks perform better than small dense networks and this relation ship holds past sparsity levels of more than 96%. The small number of weights in a Sparse WaveRNN makes it possible to sample high-fidelity audio on a mobile ph one CPU in real time. Finally, we describe a new dependency scheme for sampling that lets us trade a constant number of non-local, distant dependencies for the ability to generate samples in batches. The Batch WaveRNN produces 8 samples per step without loss of quality and offers orthogonal ways of further increasing s ampling efficiency.

****

Learning Diffusion using Hyperparameters

Dimitris Kalimeris, Yaron Singer, Karthik Subbian, Udi Weinsberg

In this paper we advocate for a hyperparametric approach to learn diffusion in t he independent cascade (IC) model. The sample complexity of this model is a func tion of the number of edges in the network and consequently learning becomes inf easible when the network is large. We study a natural restriction of the hypothe sis class using additional information available in order to dramatically reduce the sample complexity of the learning process. In particular we assume that dif fusion probabilities can be described as a function of a global hyperparameter a nd features of the individuals in the network. One of the main challenges with t his approach is that training a model reduces to optimizing a non-convex objecti ve. Despite this obstacle, we can shrink the best-known sample complexity bound for learning IC by a factor of |E|/d where |E| is the number of edges in the gra ph and d is the dimension of the hyperparameter. We show that under mild assumpt ions about the distribution generating the samples one can provably train a mode l with low generalization error. Finally, we use large-scale diffusion data from Facebook to show that a hyperparametric model using approximately 20 features p er node achieves remarkably high accuracy.

****

Signal and Noise Statistics Oblivious Orthogonal Matching Pursuit

Sreejith Kallummil, Sheetal Kalyani

Orthogonal matching pursuit (OMP) is a widely used algorithm for recovering spar se high dimensional vectors in linear regression models. The optimal performance of OMP requires a priori knowledge of either the sparsity of regression vector or noise statistics. Both these statistics are rarely known a priori and are ver y difficult to estimate. In this paper, we present a novel technique called resi dual ratio thresholding (RRT) to operate OMP without any a priori knowledge of s parsity and noise statistics and establish finite sample and large sample suppor t recovery guarantees for the same. Both analytical results and numerical simula tions in real and synthetic data sets indicate that RRT has a performance compar able to OMP with a priori knowledge of sparsity and noise statistics.

****
Residual Unfairness in Fair Machine Learning from Prejudiced Data
Nathan Kallus, Angela Zhou
Recent work in fairness in machine learning has proposed adjusting for fairness by equalizing accuracy metrics across groups and has also studied how datasets affected by historical prejudices may lead to unfair decision policies. We connect these lines of work and study the residual unfairness that arises when a fairness-adjusted predictor is not actually fair on the target population due to systematic censoring of training data by existing biased policies. This scenario is particularly common in the same applications where fairness is a concern. We characterize theoretically the impact of such censoring on standard fairness metrics for binary classifiers and provide criteria for when residual unfairness may or may not appear. We prove that, under certain conditions, fairness-adjusted classifiers will in fact induce residual unfairness that perpetuates the same injustices, against the same groups, that biased the data to begin with, thus showing that even state-of-the-art fair machine learning can have a "bias in, bias out" property. When certain benchmark data is available, we show how sample reweighting can estimate and adjust fairness metrics while accounting for censoring. We use this to study the case of Stop, Question, and Frisk (SQF) and demonstrate that attempting to adjust for fairness perpetuates the same injustices that the policy is infamous for.
****
Learn from Your Neighbor: Learning Multi-modal Mappings from Sparse Annotations
Ashwin Kalyan, Stefan Lee, Anitha Kannan, Dhruv Batra
Many structured prediction problems (particularly in vision and language domains) are ambiguous, with multiple outputs being ‘correct’ for an input {–} e.g. there are many ways of describing an image, multiple ways of translating a sentence; however, exhaustively annotating the applicability of all possible outputs is intractable due to exponentially large output spaces (e.g. all English sentences). In practice, these problems are cast as multi-class prediction, with the likelihood of only a sparse set of annotations being maximized {–} unfortunately penalizing for placing beliefs on plausible but unannotated outputs. We make and test the following hypothesis {–} for a given input, the annotations of its neighbors may serve as an additional supervisory signal. Specifically, we propose an objective that transfers supervision from neighboring examples. We first study the properties of our developed method in a controlled toy setup before reporting results on multi-label classification and two image-grounded sequence modeling tasks {–} captioning and question generation. We evaluate using standard task-specific metrics and measures of output diversity, finding consistent improvements over standard maximum likelihood training and other baselines.
****
Semi-Supervised Learning via Compact Latent Space Clustering
Konstantinos Kamnitsas, Daniel Castro, Loic Le Folgoc, Ian Walker, Ryutaro Tanno, Daniel Rueckert, Ben Glocker, Antonio Criminisi, Aditya Nori
We present a novel cost function for semi-supervised learning of neural networks that encourages compact clustering of the latent space to facilitate separation. The key idea is to dynamically create a graph over embeddings of labeled and unlabeled samples of a training batch to capture underlying structure in feature space, and use label propagation to estimate its high and low density regions. We then devise a cost function based on Markov chains on the graph that regularizes the latent space to form a single compact cluster per class, while avoiding to disturb existing clusters during optimization. We evaluate our approach on three benchmarks and compare to state-of-the art with promising results. Our approach combines the benefits of graph-based regularization with efficient, inductive inference, does not require modifications to a network architecture, and can thus be easily applied to existing networks to enable an effective use of unlabeled data.
****
Policy Optimization with Demonstrations
Bingyi Kang, Zequn Jie, Jiashi Feng

Exploration remains a significant challenge to reinforcement learning methods, e specially in environments where reward signals are sparse. Recent methods of lea rning from demonstrations have shown to be promising in overcoming exploration d ifficulties but typically require considerable high-quality demonstrations that are difficult to collect. We propose to effectively leverage available demonstra tions to guide exploration through enforcing occupancy measure matching between the learned policy and current demonstrations, and develop a novel Policy Optimi zation from Demonstration (POfD) method. We show that POfD induces implicit dyna mic reward shaping and brings provable benefits for policy improvement. Furtherm ore, it can be combined with policy gradient methods to produce state-of-the-art results, as demonstrated experimentally on a range of popular benchmark sparse-reward tasks, even when the demonstrations are few and imperfect.
****

Improving Sign Random Projections With Additional Information
Keegan Kang, Weipin Wong
Sign random projections (SRP) is a technique which allows the user to quickly es timate the angular similarity and inner products between data. We propose using additional information to improve these estimates which is easy to implement and cost efficient. We prove that the variance of our estimator is lower than the v ariance of SRP. Our proposed method can also be used together with other modific ations of SRP, such as Super-Bit LSH (SBLSH). We demonstrate the effectiveness o f our method on the MNIST test dataset and the Gisette dataset. We discuss how o ur proposed method can be extended to random projections or even other hashing a lgorithms.
****

Let’s be Honest: An Optimal No-Regret Framework for Zero-Sum Games
Ehsan Asadi Kangarshahi, Ya-Ping Hsieh, Mehmet Fatih Sahin, Volkan Cevher
We revisit the problem of solving two-player zero-sum games in the decentralized setting. We propose a simple algorithmic framework that simultaneously achieves the best rates for honest regret as well as adversarial regret, and in addition resolves the open problem of removing the logarithmic terms in convergence to t he value of the game. We achieve this goal in three steps. First, we provide a n ovel analysis of the optimistic mirror descent (OMD), showing that it can be mod ified to guarantee fast convergence for both honest regret and value of the game , when the players are playing collaboratively. Second, we propose a new algorit hm, dubbed as robust optimistic mirror descent (ROMD), which attains optimal adv ersarial regret without knowing the time horizon beforehand. Finally, we propose a simple signaling scheme, which enables us to bridge OMD and ROMD to achieve t he best of both worlds. Numerical examples are presented to support our theoreti cal claims and show that our non-adaptive ROMD algorithm can be competitive to O MD with adaptive step-size selection.
****

Continual Reinforcement Learning with Complex Synapses
Christos Kaplanis, Murray Shanahan, Claudia Clopath
Unlike humans, who are capable of continual learning over their lifetimes, artif icial neural networks have long been known to suffer from a phenomenon known as catastrophic forgetting, whereby new learning can lead to abrupt erasure of prev iously acquired knowledge. Whereas in a neural network the parameters are typica lly modelled as scalar values, an individual synapse in the brain comprises a co mplex network of interacting biochemical components that evolve at different tim escales. In this paper, we show that by equipping tabular and deep reinforcement learning agents with a synaptic model that incorporates this biological complex ity (Benna & Fusi, 2016), catastrophic forgetting can be mitigated at multiple t imescales. In particular, we find that as well as enabling continual learning ac ross sequential training of two simple tasks, it can also be used to overcome wi thin-task forgetting by reducing the need for an experience replay database.
****

LaVAN: Localized and Visible Adversarial Noise
Danny Karmon, Daniel Zoran, Yoav Goldberg
Most works on adversarial examples for deep-learning based image classifiers use

noise that, while small, covers the entire image. We explore the case where the noise is allowed to be visible but confined to a small, localized patch of the image, without covering any of the main object(s) in the image. We show that it is possible to generate localized adversarial noises that cover only 2% of the pixels in the image, none of them over the main object, and that are transferable across images and locations, and successfully fool a state-of-the-art Inception v3 model with very high success rates.
****

Riemannian Stochastic Recursive Gradient Algorithm
Hiroyuki Kasai, Hiroyuki Sato, Bamdev Mishra
Stochastic variance reduction algorithms have recently become popular for minimizing the average of a large, but finite number of loss functions on a Riemannian manifold. The present paper proposes a Riemannian stochastic recursive gradient algorithm (R-SRG), which does not require the inverse of retraction between two distant iterates on the manifold. Convergence analyses of R-SRG are performed on both retraction-convex and non-convex functions under computationally efficient retraction and vector transport operations. The key challenge is analysis of the influence of vector transport along the retraction curve. Numerical evaluations reveal that R-SRG competes well with state-of-the-art Riemannian batch and stochastic gradient algorithms.
****

Not All Samples Are Created Equal: Deep Learning with Importance Sampling
Angelos Katharopoulos, Francois Fleuret
Deep Neural Network training spends most of the computation on examples that are properly handled, and could be ignored. We propose to mitigate this phenomenon with a principled importance sampling scheme that focuses computation on "informative" examples, and reduces the variance of the stochastic gradients during training. Our contribution is twofold: first, we derive a tractable upper bound to the per-sample gradient norm, and second we derive an estimator of the variance reduction achieved with importance sampling, which enables us to switch it on when it will result in an actual speedup. The resulting scheme can be used by changing a few lines of code in a standard SGD procedure, and we demonstrate experimentally on image classification, CNN fine-tuning, and RNN training, that for a fixed wall-clock time budget, it provides a reduction of the train losses of up to an order of magnitude and a relative improvement of test errors between 5% and 17%.
****

Feasible Arm Identification
Julian Katz-Samuels, Clay Scott
We introduce the feasible arm identification problem, a pure exploration multi-armed bandit problem where the agent is given a set of D-dimensional arms and a polyhedron $P = \{x : A x \leq b \} \subset R^D$. Pulling an arm gives a random vector and the goal is to determine, using a fixed budget of T pulls, which of the arms have means belonging to P. We propose three algorithms MD-UCBE, MD-SAR, and MD-APT and provide a unified analysis establishing upper bounds for each of them. We also establish a lower bound that matches up to constants the upper bounds of MD-UCBE and MD-APT. Finally, we demonstrate the effectiveness of our algorithms on synthetic and real-world datasets.
****

Scalable Deletion-Robust Submodular Maximization: Data Summarization with Privacy and Fairness Constraints
Ehsan Kazemi, Morteza Zadimoghaddam, Amin Karbasi
Can we efficiently extract useful information from a large user-generated dataset while protecting the privacy of the users and/or ensuring fairness in representation? We cast this problem as an instance of a deletion-robust submodular maximization where part of the data may be deleted or masked due to privacy concerns or fairness criteria. We propose the first memory-efficient centralized, streaming, and distributed methods with constant-factor approximation guarantees against any number of adversarial deletions. We extensively evaluate the performance of our algorithms on real-world applications, including (i) Uber-pick up locatio

ns with location privacy constraints; (ii) feature selection with fairness const raints for income prediction and crime rate prediction; and (iii) robust to dele tion summarization of census data, consisting of 2,458,285 feature vectors. Our experiments show that our solution is robust against even $80%$ of data deletion .

****

Focused Hierarchical RNNs for Conditional Sequence Processing

Nan Rosemary Ke, Konrad ■o■na, Alessandro Sordoni, Zhouhan Lin, Adam Trischler, Yoshua Bengio, Joelle Pineau, Laurent Charlin, Christopher Pal

Recurrent Neural Networks (RNNs) with attention mechanisms have obtained state-o f-the-art results for many sequence processing tasks. Most of these models use a simple form of encoder with attention that looks over the entire sequence and a ssigns a weight to each token independently. We present a mechanism for focusing RNN encoders for sequence modelling tasks which allows them to attend to key pa rts of the input as needed. We formulate this using a multi-layer conditional hi erarchical sequence encoder that reads in one token at a time and makes a discre te decision on whether the token is relevant to the context or question being as ked. The discrete gating mechanism takes in the context embedding and the curren t hidden state as inputs and controls information flow into the layer above. We train it using policy gradient methods. We evaluate this method on several types of tasks with different attributes. First, we evaluate the method on synthetic tasks which allow us to evaluate the model for its generalization ability and pr obe the behavior of the gates in more controlled settings. We then evaluate this approach on large scale Question Answering tasks including the challenging MS M ARCO and SearchQA tasks. Our models shows consistent improvements for both tasks over prior work and our baselines. It has also shown to generalize significantl y better on synthetic tasks as compared to the baselines.

****

Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness

Michael Kearns, Seth Neel, Aaron Roth, Zhiwei Steven Wu

The most prevalent notions of fairness in machine learning fix a small collectio n of pre-defined groups (such as race or gender), and then ask for approximate p arity of some statistic of the classifier (such as false positive rate) across t hese groups. Constraints of this form are susceptible to fairness gerrymandering , in which a classifier is fair on each individual group, but badly violates the fairness constraint on structured subgroups, such as certain combinations of pr otected attribute values. We thus consider fairness across exponentially or infi nitely many subgroups, defined by a structured class of functions over the prote cted attributes. We first prove that the problem of auditing subgroup fairness f or both equality of false positive rates and statistical parity is computational ly equivalent to the problem of weak agnostic learning — which means it is hard in the worst case, even for simple structured subclasses. However, it also sugge sts that common heuristics for learning can be applied to successfully solve the auditing problem in practice. We then derive an algorithm that provably converg es in a polynomial number of steps to the best subgroup-fair distribution over c lassifiers, given access to an oracle which can solve the agnostic learning prob lem. The algorithm is based on a formulation of subgroup fairness as a zero-sum game between a Learner (the primal player) and an Auditor (the dual player). We implement a variant of this algorithm using heuristic oracles, and show that we can effectively both audit and learn fair classifiers on a real dataset.

****

Improved nearest neighbor search using auxiliary information and priority functi ons

Omid Keivani, Kaushik Sinha

Nearest neighbor search using random projection trees has recently been shown to achieve superior performance, in terms of better accuracy while retrieving less number of data points, compared to locality sensitive hashing based methods. Ho wever, to achieve acceptable nearest neighbor search accuracy for large scale ap plications, where number of data points and/or number of features can be very la rge, it requires users to maintain, store and search through large number of suc

h independent random projection trees, which may be undesirable for many practic
al applications. To address this issue, in this paper we present different searc
h strategies to improve nearest neighbor search performance of a single random p
rojection tree. Our approach exploits properties of single and multiple random p
rojections, which allows us to store meaningful auxiliary information at interna
l nodes of a random projection tree as well as to design priority functions to g
uide the search process that results in improved nearest neighbor search perform
ance. Empirical results on multiple real world datasets show that our proposed m
ethod improves the search accuracy of a single tree compared to baseline methods
.
****

ContextNet: Deep learning for Star Galaxy Classification
Noble Kennamer, David Kirkby, Alexander Ihler, Francisco Javier Sanchez-Lopez
We present a framework to compose artificial neural networks in cases where the
data cannot be treated as independent events. Our particular motivation is star
galaxy classification for ground based optical surveys. Due to a turbulent atmos
phere and imperfect instruments, a single image of an astronomical object is not
 enough to definitively classify it as a star or galaxy. Instead the context of
the surrounding objects imaged at the same time need to be considered in order t
o make an optimal classification. The model we present is divided into three dis
tinct ANNs: one designed to capture local features about each object, the second
 to compare these features across all objects in an image, and the third to make
 a final prediction for each object based on the local and compared features. By
 exploiting the ability to replicate the weights of an ANN, the model can handle
 an arbitrary and variable number of individual objects embedded in a larger exp
osure. We train and test our model on simulations of a large up and coming groun
d based survey, the Large Synoptic Survey Telescope (LSST). We compare to the st
ate of the art approach, showing improved overall performance as well as better
performance for a specific class of objects that is important for the LSST.
****

Frank-Wolfe with Subsampling Oracle
Thomas Kerdreux, Fabian Pedregosa, Alexandre d’Aspremont
We analyze two novel randomized variants of the Frank-Wolfe (FW) or conditional
gradient algorithm. While classical FW algorithms require solving a linear minim
ization problem over the domain at each iteration, the proposed method only requ
ires to solve a linear minimization problem over a small subset of the original
domain. The first algorithm that we propose is a randomized variant of the origi
nal FW algorithm and achieves a $\mathcal{O}(1/t)$ sublinear convergence rate as
 in the deterministic counterpart. The second algorithm is a randomized variant
of the Away-step FW algorithm, and again as its deterministic counterpart, reach
es linear (i.e., exponential) convergence rate making it the first provably conv
ergent randomized variant of Away-step FW. In both cases, while subsampling redu
ces the convergence rate by a constant factor, the linear minimization step can
be a fraction of the cost of that of the deterministic versions, especially when
 the data is streamed. We illustrate computational gains of both algorithms on r
egression problems, involving both ℓ_1 and latent group lasso penalties.
****

Convergence guarantees for a class of non-convex and non-smooth optimization pro
blems
Koulik Khamaru, Martin Wainwright
Non-convex optimization problems arise frequently in machine learning, including
 feature selection, structured matrix learning, mixture modeling, and neural net
work training. We consider the problem of finding critical points of a broad cla
ss of non-convex problems with non-smooth components. We analyze the behavior of
 two gradient-based methods—namely a sub-gradient method, and a proximal method.
 Our main results are to establish rates of convergence for general problems, an
d also exhibit faster rates for sub-analytic functions. As an application of our
 theory, we obtain a simplification of the popular CCCP algorithm, which retains
 all the desirable convergence properties of the original method, along with a s
ignificantly lower cost per iteration. We illustrate our methods and theory via

application to the problems of best subset selection, robust estimation, and sha
pe from shading reconstruction.
****

Fast and Scalable Bayesian Deep Learning by Weight-Perturbation in Adam
Mohammad Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, Akash Srivast
ava
Uncertainty computation in deep learning is essential to design robust and relia
ble systems. Variational inference (VI) is a promising approach for such computa
tion, but requires more effort to implement and execute compared to maximum-like
lihood methods. In this paper, we propose new natural-gradient algorithms to red
uce such efforts for Gaussian mean-field VI. Our algorithms can be implemented w
ithin the Adam optimizer by perturbing the network weights during gradient evalu
ations, and uncertainty estimates can be cheaply obtained by using the vector th
at adapts the learning rate. This requires lower memory, computation, and implem
entation effort than existing VI methods, while obtaining uncertainty estimates
of comparable quality. Our empirical results confirm this and further suggest th
at the weight-perturbation in our algorithm could be useful for exploration in r
einforcement learning and stochastic optimization.
****

Geometry Score: A Method For Comparing Generative Adversarial Networks
Valentin Khrulkov, Ivan Oseledets
One of the biggest challenges in the research of generative adversarial networks
 (GANs) is assessing the quality of generated samples and detecting various leve
ls of mode collapse. In this work, we construct a novel measure of performance o
f a GAN by comparing geometrical properties of the underlying data manifold and
the generated one, which provides both qualitative and quantitative means for ev
aluation. Our algorithm can be applied to datasets of an arbitrary nature and is
 not limited to visual data. We test the obtained metric on various real-life mo
dels and datasets and demonstrate that our method provides new insights into pro
perties of GANs.
****

Blind Justice: Fairness with Encrypted Sensitive Attributes
Niki Kilbertus, Adria Gascon, Matt Kusner, Michael Veale, Krishna Gummadi, Adria
n Weller
Recent work has explored how to train machine learning models which do not discr
iminate against any subgroup of the population as determined by sensitive attrib
utes such as gender or race. To avoid disparate treatment, sensitive attributes
should not be considered. On the other hand, in order to avoid disparate impact,
 sensitive attributes must be examined, e.g., in order to learn a fair model, or
 to check if a given model is fair. We introduce methods from secure multi-party
 computation which allow us to avoid both. By encrypting sensitive attributes, w
e show how an outcome-based fair model may be learned, checked, or have its outp
uts verified and held to account, without users revealing their sensitive attrib
utes.
****

Markov Modulated Gaussian Cox Processes for Semi-Stationary Intensity Modeling o
f Events Data
Minyoung Kim
The Cox process is a flexible event model that can account for uncertainty of th
e intensity function in the Poisson process. However, previous approaches make s
trong assumptions in terms of time stationarity, potentially failing to generali
ze when the data do not conform to the assumed stationarity conditions. In this
paper we bring up two most popular Cox models representing two extremes, and pro
pose a novel semi-stationary Cox process model that can take benefits from both
models. Our model has a set of Gaussian process latent functions governed by a l
atent stationary Markov process where we provide analytic derivations for the va
riational inference. Empirical evaluations on several synthetic and real-world e
vents data including the football shot attempts and daily earthquakes, demonstra
te that the proposed model is promising, can yield improved generalization perfo
rmance over existing approaches.

****
Disentangling by Factorising
Hyunjik Kim, Andriy Mnih
We define and address the problem of unsupervised learning of disentangled representations on data generated from independent factors of variation. We propose FactorVAE, a method that disentangles by encouraging the distribution of representations to be factorial and hence independent across the dimensions. We show that it improves upon beta-VAE by providing a better trade-off between disentanglement and reconstruction quality and being more robust to the number of training iterations. Moreover, we highlight the problems of a commonly used disentanglement metric and introduce a new metric that does not suffer from them.
****
Self-Bounded Prediction Suffix Tree via Approximate String Matching
Dongwoo Kim, Christian Walder
Prediction suffix trees (PST) provide an effective tool for sequence modelling and prediction. Current prediction techniques for PSTs rely on exact matching between the suffix of the current sequence and the previously observed sequence. We present a provably correct algorithm for learning a PST with approximate suffix matching by relaxing the exact matching condition. We then present a self-bounded enhancement of our algorithm where the depth of suffix tree grows automatically in response to the model performance on a training sequence. Through experiments on synthetic datasets as well as three real-world datasets, we show that the approximate matching PST results in better predictive performance than the other variants of PST.
****
Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)
Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, Rory sayres
The interpretation of deep learning models is a challenge due to their size, complexity, and often opaque internal state. In addition, many systems, such as image classifiers, operate on low-level features rather than high-level concepts. To address these challenges, we introduce Concept Activation Vectors (CAVs), which provide an interpretation of a neural net’s internal state in terms of human-friendly concepts. The key idea is to view the high-dimensional internal state of a neural net as an aid, not an obstacle. We show how to use CAVs as part of a technique, Testing with CAVs (TCAV), that uses directional derivatives to quantify the degree to which a user-defined concept is important to a classification result–for example, how sensitive a prediction of “zebra” is to the presence of stripes. Using the domain of image classification as a testing ground, we describe how CAVs may be used to explore hypotheses and generate insights for a standard image classification network as well as a medical application.
****
Semi-Amortized Variational Autoencoders
Yoon Kim, Sam Wiseman, Andrew Miller, David Sontag, Alexander Rush
Amortized variational inference (AVI) replaces instance-specific local inference with a global inference network. While AVI has enabled efficient training of deep generative models such as variational autoencoders (VAE), recent empirical work suggests that inference networks can produce suboptimal variational parameters. We propose a hybrid approach, to use AVI to initialize the variational parameters and run stochastic variational inference (SVI) to refine them. Crucially, the local SVI procedure is itself differentiable, so the inference network and generative model can be trained end-to-end with gradient-based optimization. This semi-amortized approach enables the use of rich generative models without experiencing the posterior-collapse phenomenon common in training VAEs for problems like text generation. Experiments show this approach outperforms strong autoregressive and variational baselines on standard text and image datasets.
****
Neural Relational Inference for Interacting Systems
Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, Richard Zemel

Interacting systems are prevalent in nature, from dynamical systems in physics t
o complex societal dynamics. The interplay of components can give rise to comple
x behavior, which can often be explained using a simple model of the system’s co
nstituent parts. In this work, we introduce the neural relational inference (NRI
) model: an unsupervised model that learns to infer interactions while simultane
ously learning the dynamics purely from observational data. Our model takes the
form of a variational auto-encoder, in which the latent code represents the unde
rlying interaction graph and the reconstruction is based on graph neural network
s. In experiments on simulated physical systems, we show that our NRI model can
accurately recover ground-truth interactions in an unsupervised manner. We furth
er demonstrate that we can find an interpretable structure and predict complex d
ynamics in real motion capture and sports tracking data.
****

An Alternative View: When Does SGD Escape Local Minima?
Bobby Kleinberg, Yuanzhi Li, Yang Yuan
Stochastic gradient descent (SGD) is widely used in machine learning. Although b
eing commonly viewed as a fast but not accurate version of gradient descent (GD)
, it always finds better solutions than GD for modern neural networks. In order
to understand this phenomenon, we take an alternative view that SGD is working o
n the convolved (thus smoothed) version of the loss function. We show that, even
 if the function f has many bad local minima or saddle points, as long as for
every point x, the weighted average of the gradients of its neighborhoods is o
ne point convex with respect to the desired solution x^*, SGD will get close t
o, and then stay around x^* with constant probability. Our result identifies a
 set of functions that SGD provably works, which is much larger than the set of
convex functions. Empirically, we observe that the loss surface of neural networ
ks enjoys nice one point convexity properties locally, therefore our theorem hel
ps explain why SGD works so well for neural networks.
****

Crowdsourcing with Arbitrary Adversaries
Matthaeus Kleindessner, Pranjal Awasthi
Most existing works on crowdsourcing assume that the workers follow the Dawid-Sk
ene model, or the one-coin model as its special case, where every worker makes m
istakes independently of other workers and with the same error probability for e
very task. We study a significant extension of this restricted model. We allow a
lmost half of the workers to deviate from the one-coin model and for those worke
rs, their probabilities of making an error to be task-dependent and to be arbitr
arily correlated. In other words, we allow for arbitrary adversaries, for which
not only error probabilities can be high, but which can also perfectly collude.
In this adversarial scenario, we design an efficient algorithm to consistently e
stimate the workers’ error probabilities.
****

Spatio-temporal Bayesian On-line Changepoint Detection with Model Selection
Jeremias Knoblauch, Theodoros Damoulas
Bayesian On-line Changepoint Detection is extended to on-line model selection an
d non-stationary spatio-temporal processes. We propose spatially structured Vect
or Autoregressions (VARs) for modelling the process between changepoints (CPs) a
nd give an upper bound on the approximation error of such models. The resulting
algorithm performs prediction, model selection and CP detection on-line. Its tim
e complexity is linear and its space complexity constant, and thus it is two ord
ers of magnitudes faster than its closest competitor. In addition, it outperform
s the state of the art for multivariate data.
****

Fast Gradient-Based Methods with Exponential Rate: A Hybrid Control Framework
Arman Sharifi Kolarijani, Peyman Mohajerin Esfahani, Tamas Keviczky
Ordinary differential equations, and in general a dynamical system viewpoint, ha
ve seen a resurgence of interest in developing fast optimization methods, mainly
 thanks to the availability of well-established analysis tools. In this study, w
e pursue a similar objective and propose a class of hybrid control systems that
adopts a 2nd-order differential equation as its continuous flow. A distinctive f

eature of the proposed differential equation in comparison with the existing lit
erature is a state-dependent, time-invariant damping term that acts as a feedbac
k control input. Given a user-defined scalar α, it is shown that the prop
osed control input steers the state trajectories to the global optimizer of a de
sired objective function with a guaranteed rate of convergence $\mathcal{O}(e^{-
\alpha t})$. Our framework requires that the objective function satisfies the so
 called Polyak–{■}ojasiewicz inequality. Furthermore, a discretization method is
 introduced such that the resulting discrete dynamical system possesses an expon
ential rate of convergence.
****

Nonconvex Optimization for Regression with Fairness Constraints
Junpei Komiyama, Akiko Takeda, Junya Honda, Hajime Shimao
The unfairness of a regressor is evaluated by measuring the correlation between
the estimator and the sensitive attribute (e.g., race, gender, age), and the coe
fficient of determination (CoD) is a natural extension of the correlation coeffi
cient when more than one sensitive attribute exists. As is well known, there is
a trade-off between fairness and accuracy of a regressor, which implies a perfec
tly fair optimizer does not always yield a useful prediction. Taking this into c
onsideration, we optimize the accuracy of the estimation subject to a user-defin
ed level of fairness. However, a fairness level as a constraint induces a noncon
vexity of the feasible region, which disables the use of an off-the-shelf convex
 optimizer. Despite such nonconvexity, we show an exact solution is available by
 using tools of global optimization theory. Furthermore, we propose a nonlinear
extension of the method by kernel representation. Unlike most of existing fairne
ss-aware machine learning methods, our method allows us to deal with numeric and
 multiple sensitive attributes.
****

On the Generalization of Equivariance and Convolution in Neural Networks to the
Action of Compact Groups
Risi Kondor, Shubhendu Trivedi
Convolutional neural networks have been extremely successful in the image recogn
ition domain because they ensure equivariance with respect to translations. Ther
e have been many recent attempts to generalize this framework to other domains,
including graphs and data lying on manifolds. In this paper we give a rigorous,
theoretical treatment of convolution and equivariance in neural networks with re
spect to not just translations, but the action of any compact group. Our main re
sult is to prove that (given some natural constraints) convolutional structure i
s not just a sufficient, but also a necessary condition for equivariance to the
action of a compact group. Our exposition makes use of concepts from representat
ion theory and noncommutative harmonic analysis and derives new generalized conv
olution formulae.
****

Compiling Combinatorial Prediction Games
Frederic Koriche
In online optimization, the goal is to iteratively choose solutions from a decis
ion space, so as to minimize the average cost over time. As long as this decisio
n space is described by combinatorial constraints, the problem is generally intr
actable. In this paper, we consider the paradigm of compiling the set of combina
torial constraints into a deterministic and Decomposable Negation Normal Form (d
DNNF) circuit, for which the tasks of linear optimization and solution sampling
take linear time. Based on this framework, we provide efficient characterization
s of existing combinatorial prediction strategies, with a particular attention t
o mirror descent techniques. These strategies are compared on several real-world
 benchmarks for which the set of Boolean constraints is preliminarily compiled i
nto a dDNNF circuit.
****

Dynamic Evaluation of Neural Sequence Models
Ben Krause, Emmanuel Kahembwe, Iain Murray, Steve Renals
We explore dynamic evaluation, where sequence models are adapted to the recent s
equence history using gradient descent, assigning higher probabilities to re-occ

urring sequential patterns. We develop a dynamic evaluation approach that outper forms existing adaptation approaches in our comparisons. We apply dynamic evalua tion to outperform all previous word-level perplexities on the Penn Treebank and WikiText-2 datasets (achieving 51.1 and 44.3 respectively) and all previous cha racter-level cross-entropies on the text8 and Hutter Prize datasets (achieving 1 .19 bits/char and 1.08 bits/char respectively).
****

Semiparametric Contextual Bandits
Akshay Krishnamurthy, Zhiwei Steven Wu, Vasilis Syrgkanis
This paper studies semiparametric contextual bandits, a generalization of the li near stochastic bandit problem where the reward for a chosen action is modeled a s a linear function of known action features confounded by a non-linear action-i ndependent term. We design new algorithms that achieve $\tilde{O}(d\sqrt{T})$ re gret over T rounds, when the linear function is d-dimensional, which matches the best known bounds for the simpler unconfounded case and improves on a recen t result of Greenwald et al. (2017). Via an empirical evaluation, we show that o ur algorithms outperform prior approaches when there are non-linear confounding effects on the rewards. Technically, our algorithms use a new reward estimator i nspired by doubly-robust approaches and our proofs require new concentration ine qualities for self-normalized martingales.
****

Fast Maximization of Non-Submodular, Monotonic Functions on the Integer Lattice
Alan Kuhnle, J. David Smith, Victoria Crawford, My Thai
The optimization of submodular functions on the integer lattice has received muc h attention recently, but the objective functions of many applications are non-s ubmodular. We provide two approximation algorithms for maximizing a non-submodul ar function on the integer lattice subject to a cardinality constraint; these ar e the first algorithms for this purpose that have polynomial query complexity. W e propose a general framework for influence maximization on the integer lattice that generalizes prior works on this topic, and we demonstrate the efficiency of our algorithms in this context.
****

Accurate Uncertainties for Deep Learning Using Calibrated Regression
Volodymyr Kuleshov, Nathan Fenner, Stefano Ermon
Methods for reasoning under uncertainty are a key building block of accurate and reliable machine learning systems. Bayesian methods provide a general framework to quantify uncertainty. However, because of model misspecification and the use of approximate inference, Bayesian uncertainty estimates are often inaccurate {—} for example, a 90% credible interval may not contain the true outcome 90% of the time. Here, we propose a simple procedure for calibrating any regression alg orithm; when applied to Bayesian and probabilistic models, it is guaranteed to p roduce calibrated uncertainty estimates given enough data. Our procedure is insp ired by Platt scaling and extends previous work on classification. We evaluate t his approach on Bayesian linear regression, feedforward, and recurrent neural ne tworks, and find that it consistently outputs well-calibrated credible intervals while improving performance on time series forecasting and model-based reinforc ement learning tasks.
****

Trainable Calibration Measures for Neural Networks from Kernel Mean Embeddings
Aviral Kumar, Sunita Sarawagi, Ujjwal Jain
Modern neural networks have recently been found to be poorly calibrated, primari ly in the direction of over-confidence. Methods like entropy penalty and tempera ture smoothing improve calibration by clamping confidence, but in doing so compr omise the many legitimately confident predictions. We propose a more principled fix that minimizes an explicit calibration error during training. We present MMC E, a RKHS kernel based measure of calibration that is efficiently trainable alon gside the negative likelihood loss without careful hyper-parameter tuning. Theor etically too, MMCE is a sound measure of calibration that is minimized at perfec t calibration, and whose finite sample estimates are consistent and enjoy fast c onvergence rates. Extensive experiments on several network architectures demonst

rate that MMCE is a fast, stable, and accurate method to minimize calibration er
ror while maximally preserving the number of high confidence predictions.
****
Data-Dependent Stability of Stochastic Gradient Descent
Ilja Kuzborskij, Christoph Lampert
We establish a data-dependent notion of algorithmic stability for Stochastic Gra
dient Descent (SGD), and employ it to develop novel generalization bounds. This
is in contrast to previous distribution-free algorithmic stability results for S
GD which depend on the worst-case constants. By virtue of the data-dependent arg
ument, our bounds provide new insights into learning with SGD on convex and non-
convex problems. In the convex case, we show that the bound on the generalizatio
n error depends on the risk at the initialization point. In the non-convex case,
 we prove that the expected curvature of the objective function around the initi
alization point has crucial influence on the generalization error. In both cases
, our results suggest a simple data-driven strategy to stabilize SGD by pre-scre
ening its initialization. As a corollary, our results allow us to show optimisti
c generalization bounds that exhibit fast convergence rates for SGD subject to a
 vanishing empirical risk and low noise of stochastic gradient.
****
Explicit Inductive Bias for Transfer Learning with Convolutional Networks
Xuhong LI, Yves Grandvalet, Franck Davoine
In inductive transfer learning, fine-tuning pre-trained convolutional networks s
ubstantially outperforms training from scratch. When using fine-tuning, the unde
rlying assumption is that the pre-trained model extracts generic features, which
 are at least partially relevant for solving the target task, but would be diffi
cult to extract from the limited amount of data available on the target task. Ho
wever, besides the initialization with the pre-trained model and the early stopp
ing, there is no mechanism in fine-tuning for retaining the features learned on
the source task. In this paper, we investigate several regularization schemes th
at explicitly promote the similarity of the final solution with the initial mode
l. We show the benefit of having an explicit inductive bias towards the initial
model, and we eventually recommend a simple L^2 penalty with the pre-trained m
odel being a reference as the baseline of penalty for transfer learning tasks.
****
Understanding the Loss Surface of Neural Networks for Binary Classification
SHIYU LIANG, Ruoyu Sun, Yixuan Li, Rayadurgam Srikant
It is widely conjectured that training algorithms for neural networks are succes
sful because all local minima lead to similar performance; for example, see (LeC
un et al., 2015; Choromanska et al., 2015; Dauphin et al., 2014). Performance is
 typically measured in terms of two metrics: training performance and generaliza
tion performance. Here we focus on the training performance of neural networks f
or binary classification, and provide conditions under which the training error
is zero at all local minima of appropriately chosen surrogate loss functions. Ou
r conditions are roughly in the following form: the neurons have to be increasin
g and strictly convex, the neural network should either be single-layered or is
multi-layered with a shortcut-like connection, and the surrogate loss function s
hould be a smooth version of hinge loss. We also provide counterexamples to show
 that, when these conditions are relaxed, the result may not hold.
****
Mixed batches and symmetric discriminators for GAN training
Thomas LUCAS, Corentin Tallec, Yann Ollivier, Jakob Verbeek
Generative adversarial networks (GANs) are pow- erful generative models based on
 providing feed- back to a generative network via a discriminator network. Howev
er, the discriminator usually as- sesses individual samples. This prevents the d
is- criminator from accessing global distributional statistics of generated samp
les, and often leads to mode dropping: the generator models only part of the tar
get distribution. We propose to feed the discriminator with mixed batches of tru
e and fake samples, and train it to predict the ratio of true samples in the bat
ch. The latter score does not depend on the order of samples in a batch. Rather
than learning this invariance, we introduce a generic permutation-invariant disc

riminator ar- chitecture. This architecture is provably a uni- versal approximat or of all symmetric functions. Experimentally, our approach reduces mode col- la pse in GANs on two synthetic datasets, and obtains good results on the CIFAR10 a nd CelebA datasets, both qualitatively and quantitatively.
****

Binary Partitions with Approximate Minimum Impurity
Eduardo Laber, Marco Molinaro, Felipe Mello Pereira
The problem of splitting attributes is one of the main steps in the construction of decision trees. In order to decide the best split, impurity measures such as Entropy and Gini are widely used. In practice, decision-tree inducers use heuri stics for finding splits with small impurity when they consider nominal attribut es with a large number of distinct values. However, there are no known guarantee s for the quality of the splits obtained by these heuristics. To fill this gap, we propose two new splitting procedures that provably achieve near-optimal impur ity. We also report experiments that provide evidence that the proposed methods are interesting candidates to be employed in splitting nominal attributes with m any values during decision tree/random forest induction.
****

Canonical Tensor Decomposition for Knowledge Base Completion
Timothee Lacroix, Nicolas Usunier, Guillaume Obozinski
The problem of Knowledge Base Completion can be framed as a 3rd-order binary ten sor completion problem. In this light, the Canonical Tensor Decomposition (CP) s eems like a natural solution; however, current implementations of CP on standard Knowledge Base Completion benchmarks are lagging behind their competitors. In t his work, we attempt to understand the limits of CP for knowledge base completio n. First, we motivate and test a novel regularizer, based on tensor nuclear p-no rms. Then, we present a reformulation of the problem that makes it invariant to arbitrary choices in the inclusion of predicates or their reciprocals in the dat aset. These two methods combined allow us to beat the current state of the art o n several datasets with a CP decomposition, and obtain even better results using the more advanced ComplEx model.
****

Generalization without Systematicity: On the Compositional Skills of Sequence-to -Sequence Recurrent Networks
Brenden Lake, Marco Baroni
Humans can understand and produce new utterances effortlessly, thanks to their c ompositional skills. Once a person learns the meaning of a new verb "dax," he or she can immediately understand the meaning of "dax twice" or "sing and dax." In this paper, we introduce the SCAN domain, consisting of a set of simple composi tional navigation commands paired with the corresponding action sequences. We th en test the zero-shot generalization capabilities of a variety of recurrent neur al networks (RNNs) trained on SCAN with sequence-to-sequence methods. We find th at RNNs can make successful zero-shot generalizations when the differences betwe en training and test commands are small, so that they can apply "mix-and-match" strategies to solve the task. However, when generalization requires systematic c ompositional skills (as in the "dax" example above), RNNs fail spectacularly. We conclude with a proof-of-concept experiment in neural machine translation, sugg esting that lack of systematicity might be partially responsible for neural netw orks’ notorious training data thirst.
****

An Estimation and Analysis Framework for the Rasch Model
Andrew Lan, Mung Chiang, Christoph Studer
The Rasch model is widely used for item response analysis in applications rangin g from recommender systems to psychology, education, and finance. While a number of estimators have been proposed for the Rasch model over the last decades, the associated analytical performance guarantees are mostly asymptotic. This paper provides a framework that relies on a novel linear minimum mean-squared error (L -MMSE) estimator which enables an exact, nonasymptotic, and closed-form analysis of the parameter estimation error under the Rasch model. The proposed framework provides guidelines on the number of items and responses required to attain low

estimation errors in tests or surveys. We furthermore demonstrate its efficacy on a number of real-world collaborative filtering datasets, which reveals that the proposed L-MMSE estimator performs on par with state-of-the-art nonlinear estimators in terms of predictive performance.
****
Partial Optimality and Fast Lower Bounds for Weighted Correlation Clustering
Jan-Hendrik Lange, Andreas Karrenbauer, Bjoern Andres
Weighted correlation clustering is hard to solve and hard to approximate for general graphs. Its applications in network analysis and computer vision call for efficient algorithms. To this end, we make three contributions: We establish partial optimality conditions that can be checked efficiently, and doing so recursively solves the problem for series-parallel graphs to optimality, in linear time. We exploit the packing dual of the problem to compute a heuristic, but non-trivial lower bound faster than that of a canonical linear program relaxation. We introduce a re-weighting with the dual solution by which efficient local search algorithms converge to better feasible solutions. The effectiveness of our methods is demonstrated empirically on a number of benchmark instances.
****
Deep Linear Networks with Arbitrary Loss: All Local Minima Are Global
Thomas Laurent, James Brecht
We consider deep linear networks with arbitrary convex differentiable loss. We provide a short and elementary proof of the fact that all local minima are global minima if the hidden layers are either 1) at least as wide as the input layer, or 2) at least as wide as the output layer. This result is the strongest possible in the following sense: If the loss is convex and Lipschitz but not differentiable then deep linear networks can have sub-optimal local minima.
****
The Multilinear Structure of ReLU Networks
Thomas Laurent, James Brecht
We study the loss surface of neural networks equipped with a hinge loss criterion and ReLU or leaky ReLU nonlinearities. Any such network defines a piecewise multilinear form in parameter space. By appealing to harmonic analysis we show that all local minima of such network are non-differentiable, except for those minima that occur in a region of parameter space where the loss surface is perfectly flat. Non-differentiable minima are therefore not technicalities or pathologies; they are heart of the problem when investigating the loss of ReLU networks. As a consequence, we must employ techniques from nonsmooth analysis to study these loss surfaces. We show how to apply these techniques in some illustrative cases.
****
Hierarchical Imitation and Reinforcement Learning
Hoang Le, Nan Jiang, Alekh Agarwal, Miroslav Dudik, Yisong Yue, Hal Daumé III
We study how to effectively leverage expert feedback to learn sequential decision-making policies. We focus on problems with sparse rewards and long time horizons, which typically pose significant challenges in reinforcement learning. We propose an algorithmic framework, called hierarchical guidance, that leverages the hierarchical structure of the underlying problem to integrate different modes of expert interaction. Our framework can incorporate different combinations of imitation learning (IL) and reinforcement learning (RL) at different levels, leading to dramatic reductions in both expert effort and cost of exploration. Using long-horizon benchmarks, including Montezuma’s Revenge, we demonstrate that our approach can learn significantly faster than hierarchical RL, and be significantly more label-efficient than standard IL. We also theoretically analyze labeling cost for certain instantiations of our framework.
****
Gradient-Based Meta-Learning with Learned Layerwise Metric and Subspace
Yoonho Lee, Seungjin Choi
Gradient-based meta-learning methods leverage gradient descent to learn the commonalities among various tasks. While previous such methods have been successful in meta-learning tasks, they resort to simple gradient descent during meta-testi

ng. Our primary contribution is the MT-net, which enables the meta-learner to le
arn on each layer’s activation space a subspace that the task-specific learner p
erforms gradient descent on. Additionally, a task-specific learner of an MT-net
performs gradient descent with respect to a meta-learned distance metric, which
warps the activation space to be more sensitive to task identity. We demonstrate
 that the dimension of this learned subspace reflects the complexity of the task
-specific learner’s adaptation task, and also that our model is less sensitive t
o the choice of initial learning rates than previous gradient-based meta-learnin
g methods. Our method achieves state-of-the-art or comparable performance on few
-shot classification and regression tasks.
****

Deep Reinforcement Learning in Continuous Action Spaces: a Case Study in the Gam
e of Simulated Curling

Kyowoon Lee, Sol-A Kim, Jaesik Choi, Seong-Whan Lee

Many real-world applications of reinforcement learning require an agent to selec
t optimal actions from continuous spaces. Recently, deep neural networks have su
ccessfully been applied to games with discrete actions spaces. However, deep neu
ral networks for discrete actions are not suitable for devising strategies for g
ames where a very small change in an action can dramatically affect the outcome.
 In this paper, we present a new self-play reinforcement learning framework whic
h equips a continuous search algorithm which enables to search in continuous act
ion spaces with a kernel regression method. Without any hand-crafted features, o
ur network is trained by supervised learning followed by self-play reinforcement
 learning with a high-fidelity simulator for the Olympic sport of curling. The p
rogram trained under our framework outperforms existing programs equipped with s
everal hand-crafted features and won an international digital curling competitio
n.
****

Gated Path Planning Networks

Lisa Lee, Emilio Parisotto, Devendra Singh Chaplot, Eric Xing, Ruslan Salakhutdi
nov

Value Iteration Networks (VINs) are effective differentiable path planning modul
es that can be used by agents to perform navigation while still maintaining end-
to-end differentiability of the entire architecture. Despite their effectiveness
, they suffer from several disadvantages including training instability, random
seed sensitivity, and other optimization problems. In this work, we reframe VINs
 as recurrent-convolutional networks which demonstrates that VINs couple recurre
nt convolutions with an unconventional max-pooling activation. From this perspec
tive, we argue that standard gated recurrent update equations could potentially
alleviate the optimization issues plaguing VIN. The resulting architecture, whic
h we call the Gated Path Planning Network, is shown to empirically outperform VI
N on a variety of metrics such as learning speed, hyperparameter sensitivity, it
eration count, and even generalization. Furthermore, we show that this performan
ce gap is consistent across different maze transition types, maze sizes and even
 show success on a challenging 3D environment, where the planner is only provide
d with first-person RGB images.
****

Deep Asymmetric Multi-task Feature Learning

Hae Beom Lee, Eunho Yang, Sung Ju Hwang

We propose Deep Asymmetric Multitask Feature Learning (Deep-AMTFL) which can lea
rn deep representations shared across multiple tasks while effectively preventin
g negative transfer that may happen in the feature sharing process. Specifically
, we introduce an asymmetric autoencoder term that allows reliable predictors fo
r the easy tasks to have high contribution to the feature learning while suppres
sing the influences of unreliable predictors for more difficult tasks. This allo
ws the learning of less noisy representations, and enables unreliable predictors
 to exploit knowledge from the reliable predictors via the shared latent feature
s. Such asymmetric knowledge transfer through shared features is also more scala
ble and efficient than inter-task asymmetric transfer. We validate our Deep-AMTF
L model on multiple benchmark datasets for multitask learning and image classifi

cation, on which it significantly outperforms existing symmetric and asymmetric multitask learning models, by effectively preventing negative transfer in deep feature learning.
****

Noise2Noise: Learning Image Restoration without Clean Data
Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, Timo Aila
We apply basic statistical reasoning to signal reconstruction by machine learning - learning to map corrupted observations to clean signals - with a simple and powerful conclusion: it is possible to learn to restore images by only looking at corrupted examples, at performance at and sometimes exceeding training using clean data, without explicit image priors or likelihood models of the corruption. In practice, we show that a single model learns photographic noise removal, denoising synthetic Monte Carlo images, and reconstruction of undersampled MRI scans - all corrupted by different processes - based on noisy data only.
****

Out-of-sample extension of graph adjacency spectral embedding
Keith Levin, Fred Roosta, Michael Mahoney, Carey Priebe
Many popular dimensionality reduction procedures have out-of-sample extensions, which allow a practitioner to apply a learned embedding to observations not seen in the initial training sample. In this work, we consider the problem of obtaining an out-of-sample extension for the adjacency spectral embedding, a procedure for embedding the vertices of a graph into Euclidean space. We present two different approaches to this problem, one based on a least-squares objective and the other based on a maximum-likelihood formulation. We show that if the graph of interest is drawn according to a certain latent position model called a random dot product graph, then both of these out-of-sample extensions estimate the true latent position of the out-of-sample vertex with the same error rate. Further, we prove a central limit theorem for the least-squares-based extension, showing that the estimate is asymptotically normal about the truth in the large-graph limit.
****

An Optimal Control Approach to Deep Learning and Applications to Discrete-Weight Neural Networks
Qianxiao Li, Shuji Hao
Deep learning is formulated as a discrete-time optimal control problem. This allows one to characterize necessary conditions for optimality and develop training algorithms that do not rely on gradients with respect to the trainable parameters. In particular, we introduce the discrete-time method of successive approximations (MSA), which is based on the Pontryagin’s maximum principle, for training neural networks. A rigorous error estimate for the discrete MSA is obtained, which sheds light on its dynamics and the means to stabilize the algorithm. The developed methods are applied to train, in a rather principled way, neural networks with weights that are constrained to take values in a discrete set. We obtain competitive performance and interestingly, very sparse weights in the case of ternary networks, which may be useful in model deployment in low-memory devices.
****

Towards Binary-Valued Gates for Robust LSTM Training
Zhuohan Li, Di He, Fei Tian, Wei Chen, Tao Qin, Liwei Wang, Tieyan Liu
Long Short-Term Memory (LSTM) is one of the most widely used recurrent structures in sequence modeling. It aims to use gates to control information flow (e.g., whether to skip some information or not) in the recurrent computations, although its practical implementation based on soft gates only partially achieves this goal. In this paper, we propose a new way for LSTM training, which pushes the output values of the gates towards 0 or 1. By doing so, we can better control the information flow: the gates are mostly open or closed, instead of in a middle state, which makes the results more interpretable. Empirical studies show that (1) Although it seems that we restrict the model capacity, there is no performance drop: we achieve better or comparable performances due to its better generalization ability; (2) The outputs of gates are not sensitive to their inputs: we can e

asily compress the LSTM unit in multiple ways, e.g., low-rank approximation and low-precision approximation. The compressed models are even better than the base line models without compression.
****

On the Limitations of First-Order Approximation in GAN Dynamics
Jerry Li, Aleksander Madry, John Peebles, Ludwig Schmidt
While Generative Adversarial Networks (GANs) have demonstrated promising perform ance on multiple vision tasks, their learning dynamics are not yet well understo od, both in theory and in practice. To address this issue, we study GAN dynamics in a simple yet rich parametric model that exhibits several of the common probl ematic convergence behaviors such as vanishing gradients, mode collapse, and div erging or oscillatory behavior. In spite of the non-convex nature of our model, we are able to perform a rigorous theoretical analysis of its convergence behavi or. Our analysis reveals an interesting dichotomy: a GAN with an optimal discrim inator provably converges, while first order approximations of the discriminator steps lead to unstable GAN dynamics and mode collapse. Our result suggests that using first order discriminator steps (the de-facto standard in most existing G AN setups) might be one of the factors that makes GAN training challenging in pr actice.
****

Submodular Hypergraphs: p-Laplacians, Cheeger Inequalities and Spectral Clusteri ng
Pan Li, Olgica Milenkovic
We introduce submodular hypergraphs, a family of hypergraphs that have different submodular weights associated with different cuts of hyperedges. Submodular hyp ergraphs arise in cluster- ing applications in which higher-order structures car ry relevant information. For such hypergraphs, we define the notion of p-Laplaci ans and derive corresponding nodal domain theorems and k-way Cheeger inequalitie s. We conclude with the description of algorithms for computing the spectra of 1 - and 2-Laplacians that constitute the basis of new spectral hypergraph clusteri ng methods.
****

The Well-Tempered Lasso
Yuanzhi Li, Yoram Singer
We study the complexity of the entire regularization path for least squares regr ession with 1-norm penalty, known as the Lasso. Every regression parameter in th e Lasso changes linearly as a function of the regularization value. The number o f changes is regarded as the Lasso’s complexity. Experimental results using exac t path following exhibit polynomial complexity of the Lasso in the problem size. Alas, the path complexity of the Lasso on artificially designed regression prob lems is exponential We use smoothed analysis as a mechanism for bridging the gap between worst case settings and the de facto low complexity. Our analysis assum es that the observed data has a tiny amount of intrinsic noise. We then prove th at the Lasso’s complexity is polynomial in the problem size.
****

Estimation of Markov Chain via Rank-Constrained Likelihood
Xudong Li, Mengdi Wang, Anru Zhang
This paper studies the estimation of low-rank Markov chains from empirical traje ctories. We propose a non-convex estimator based on rank-constrained likelihood maximization. Statistical upper bounds are provided for the Kullback-Leiber dive rgence and the ℓ_2 risk between the estimator and the true transition matri x. The estimator reveals a compressed state space of the Markov chain. We also d evelop a novel DC (difference of convex function) programming algorithm to tackl e the rank-constrained non-smooth optimization problem. Convergence results are established. Experiments show that the proposed estimator achieves better empiri cal performance than other popular approaches.
****

Asynchronous Decentralized Parallel Stochastic Gradient Descent
Xiangru Lian, Wei Zhang, Ce Zhang, Ji Liu
Most commonly used distributed machine learning systems are either synchronous o

r centralized asynchronous. Synchronous algorithms like AllReduce-SGD perform po orly in a heterogeneous environment, while asynchronous algorithms using a param eter server suffer from 1) communication bottleneck at parameter servers when wo rkers are many, and 2) significantly worse convergence when the traffic to param eter server is congested. Can we design an algorithm that is robust in a heterog eneous environment, while being communication efficient and maintaining the best -possible convergence rate? In this paper, we propose an asynchronous decentrali zed stochastic gradient decent algorithm (AD-PSGD) satisfying all above expectat ions. Our theoretical analysis shows AD-PSGD converges at the optimal $O(1/\sqrt {K})$ rate as SGD and has linear speedup w.r.t. number of workers. Empirically, AD-PSGD outperforms the best of decentralized parallel SGD (D-PSGD), asynchronou s parallel SGD (A-PSGD), and standard data parallel SGD (AllReduce-SGD), often b y orders of magnitude in a heterogeneous environment. When training ResNet-50 on ImageNet with up to 128 GPUs, AD-PSGD converges (w.r.t epochs) similarly to the AllReduce-SGD, but each epoch can be up to 4-8x faster than its synchronous cou nterparts in a network-sharing HPC environment. To the best of our knowledge, AD -PSGD is the first asynchronous algorithm that achieves a similar epoch-wise con vergence rate as AllReduce-SGD, at an over 100-GPU scale.

****

RLlib: Abstractions for Distributed Reinforcement Learning

Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Ken Goldber g, Joseph Gonzalez, Michael Jordan, Ion Stoica

Reinforcement learning (RL) algorithms involve the deep nesting of highly irregu lar computation patterns, each of which typically exhibits opportunities for dis tributed computation. We argue for distributing RL components in a composable wa y by adapting algorithms for top-down hierarchical control, thereby encapsulatin g parallelism and resource requirements within short-running compute tasks. We d emonstrate the benefits of this principle through RLlib: a library that provides scalable software primitives for RL. These primitives enable a broad range of a lgorithms to be implemented with high performance, scalability, and substantial code reuse. RLlib is available as part of the open source Ray project at http:// rllib.io/.

****

On the Spectrum of Random Features Maps of High Dimensional Data

Zhenyu Liao, Romain Couillet

Random feature maps are ubiquitous in modern statistical machine learning, where they generalize random projections by means of powerful, yet often difficult to analyze nonlinear operators. In this paper we leverage the "concentration" phen omenon induced by random matrix theory to perform a spectral analysis on the Gra m matrix of these random feature maps, here for Gaussian mixture models of simul taneously large dimension and size. Our results are instrumental to a deeper und erstanding on the interplay of the nonlinearity and the statistics of the data, thereby allowing for a better tuning of random feature-based techniques.

****

The Dynamics of Learning: A Random Matrix Approach

Zhenyu Liao, Romain Couillet

Understanding the learning dynamics of neural networks is one of the key issues for the improvement of optimization algorithms as well as for the theoretical co mprehension of why deep neural nets work so well today. In this paper, we introd uce a random matrix-based framework to analyze the learning dynamics of a single -layer linear network on a binary classification problem, for data of simultaneo usly large dimension and size, trained by gradient descent. Our results provide rich insights into common questions in neural nets, such as overfitting, early s topping and the initialization of training, thereby opening the door for future studies of more elaborate structures and models appearing in today’s neural netw orks.

****

Reviving and Improving Recurrent Back-Propagation

Renjie Liao, Yuwen Xiong, Ethan Fetaya, Lisa Zhang, KiJung Yoon, Xaq Pitkow, Raq uel Urtasun, Richard Zemel

In this paper, we revisit the recurrent back-propagation (RBP) algorithm, discuss the conditions under which it applies as well as how to satisfy them in deep neural networks. We show that RBP can be unstable and propose two variants based on conjugate gradient on the normal equations (CG-RBP) and Neumann series (Neumann-RBP). We further investigate the relationship between Neumann-RBP and back propagation through time (BPTT) and its truncated version (TBPTT). Our Neumann-RBP has the same time complexity as TBPTT but only requires constant memory, whereas TBPTT's memory cost scales linearly with the number of truncation steps. We examine all RBP variants along with BPTT and TBPTT in three different application domains: associative memory with continuous Hopfield networks, document classification in citation networks using graph neural networks and hyperparameter optimization for fully connected networks. All experiments demonstrate that RBPs, especially the Neumann-RBP variant, are efficient and effective for optimizing convergent recurrent neural networks.

****

Optimal Distributed Learning with Multi-pass Stochastic Gradient Methods

Junhong Lin, Volkan Cevher

We study generalization properties of distributed algorithms in the setting of nonparametric regression over a reproducing kernel Hilbert space (RKHS). We investigate distributed stochastic gradient methods (SGM), with mini-batches and multi-passes over the data. We show that optimal generalization error bounds can be retained for distributed SGM provided that the partition level is not too large. Our results are superior to the state-of-the-art theory, covering the cases that the regression function may not be in the hypothesis spaces. Particularly, our results show that distributed SGM has a smaller theoretical computational complexity, compared with distributed kernel ridge regression (KRR) and classic SGM.

****

Optimal Rates of Sketched-regularized Algorithms for Least-Squares Regression over Hilbert Spaces

Junhong Lin, Volkan Cevher

We investigate regularized algorithms combining with projection for least-squares regression problem over a Hilbert space, covering nonparametric regression over a reproducing kernel Hilbert space. We prove convergence results with respect to variants of norms, under a capacity assumption on the hypothesis space and a regularity condition on the target function. As a result, we obtain optimal rates for regularized algorithms with randomized sketches, provided that the sketch dimension is proportional to the effective dimension up to a logarithmic factor. As a byproduct, we obtain similar results for Nyström regularized algorithms. Our results provide optimal, distribution-dependent rates for sketched/Nyström regularized algorithms, considering both the attainable and non-attainable cases.

****

Level-Set Methods for Finite-Sum Constrained Convex Optimization

Qihang Lin, Runchao Ma, Tianbao Yang

We consider the constrained optimization where the objective function and the constraints are defined as summation of finitely many loss functions. This model has applications in machine learning such as Neyman-Pearson classification. We consider two level-set methods to solve this class of problems, an existing inexact Newton method and a new feasible level-set method. To update the level parameter towards the optimality, both methods require an oracle that generates upper and lower bounds as well as an affine-minorant of the level function. To construct the desired oracle, we reformulate the level function as the value of a saddle-point problem using the conjugate and perspective of the loss functions. Then a stochastic variance-reduced gradient method with a special Bregman divergence is proposed as the oracle for solving that saddle-point problem. The special divergence ensures the proximal mapping in each iteration can be solved in a closed form. The total complexity of both level-set methods using the proposed oracle are analyzed.

****

Detecting and Correcting for Label Shift with Black Box Predictors

Zachary Lipton, Yu-Xiang Wang, Alexander Smola

Faced with distribution shift between training and test set, we wish to detect and quantify the shift, and to correct our classifiers without test set labels. Motivated by medical diagnosis, where diseases (targets), cause symptoms (observations), we focus on label shift, where the label marginal p(y) changes but the conditional p(x| y) does not. We propose Black Box Shift Estimation (BBSE) to estimate the test distribution p(y). BBSE exploits arbitrary black box predictors to reduce dimensionality prior to shift correction. While better predictors give tighter estimates, BBSE works even when predictors are biased, inaccurate, or uncalibrated, so long as their confusion matrices are invertible. We prove BBSE's consistency, bound its error, and introduce a statistical test that uses BBSE to detect shift. We also leverage BBSE to correct classifiers. Experiments demonstrate accurate estimates and improved prediction, even on high-dimensional datasets of natural images.

****

Generalized Robust Bayesian Committee Machine for Large-scale Gaussian Process Regression

Haitao Liu, Jianfei Cai, Yi Wang, Yew Soon Ong

In order to scale standard Gaussian process (GP) regression to large-scale datasets, aggregation models employ factorized training process and then combine predictions from distributed experts. The state-of-the-art aggregation models, however, either provide inconsistent predictions or require time-consuming aggregation process. We first prove the inconsistency of typical aggregations using disjoint or random data partition, and then present a consistent yet efficient aggregation model for large-scale GP. The proposed model inherits the advantages of aggregations, e.g., closed-form inference and aggregation, parallelization and distributed computing. Furthermore, theoretical and empirical analyses reveal that the new aggregation model performs better due to the consistent predictions that converge to the true underlying function when the training size approaches infinity.

****

Towards Black-box Iterative Machine Teaching

Weiyang Liu, Bo Dai, Xingguo Li, Zhen Liu, James Rehg, Le Song

In this paper, we make an important step towards the black-box machine teaching by considering the cross-space machine teaching, where the teacher and the learner use different feature representations and the teacher can not fully observe the learner's model. In such scenario, we study how the teacher is still able to teach the learner to achieve faster convergence rate than the traditional passive learning. We propose an active teacher model that can actively query the learner (i.e., make the learner take exams) for estimating the learner's status and provably guide the learner to achieve faster convergence. The sample complexities for both teaching and query are provided. In the experiments, we compare the proposed active teacher with the omniscient teacher and verify the effectiveness of the active teacher model.

****

Delayed Impact of Fair Machine Learning

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, Moritz Hardt

Fairness in machine learning has predominantly been studied in static classification settings without concern for how decisions change the underlying population over time. Conventional wisdom suggests that fairness criteria promote the long-term well-being of those groups they aim to protect. We study how static fairness criteria interact with temporal indicators of well-being, such as long-term improvement, stagnation, and decline in a variable of interest. We demonstrate that even in a one-step feedback model, common fairness criteria in general do not promote improvement over time, and may in fact cause harm in cases where an unconstrained objective would not. We completely characterize the delayed impact of three standard criteria, contrasting the regimes in which these exhibit qualitatively different behavior. In addition, we find that a natural form of measurement error broadens the regime in which fairness criteria perform favorably. Our results highlight the importance of measurement and temporal modeling in the evaluation of fairness criteria, suggesting a range of new challenges and trade-offs

.
****
A Two-Step Computation of the Exact GAN Wasserstein Distance
Huidong Liu, Xianfeng GU, Dimitris Samaras
In this paper, we propose a two-step method to compute the Wasserstein distance in Wasserstein Generative Adversarial Networks (WGANs): 1) The convex part of our objective can be solved by linear programming; 2) The non-convex residual can be approximated by a deep neural network. We theoretically prove that the proposed formulation is equivalent to the discrete Monge-Kantorovich dual formulation. Furthermore, we give the approximation error bound of the Wasserstein distance and the error bound of generalizing the Wasserstein distance from discrete to continuous distributions. Our approach optimizes the exact Wasserstein distance, obviating the need for weight clipping previously used in WGANs. Results on synthetic data show that the our method computes the Wasserstein distance more accurately. Qualitative and quantitative results on MNIST, LSUN and CIFAR-10 datasets show that the proposed method is more efficient than state-of-the-art WGAN methods, and still produces images of comparable quality.
****
Open Category Detection with PAC Guarantees
Si Liu, Risheek Garrepalli, Thomas Dietterich, Alan Fern, Dan Hendrycks
Open category detection is the problem of detecting "alien" test instances that belong to categories or classes that were not present in the training data. In many applications, reliably detecting such aliens is central to ensuring the safety and accuracy of test set predictions. Unfortunately, there are no algorithms that provide theoretical guarantees on their ability to detect aliens under general assumptions. Further, while there are algorithms for open category detection, there are few empirical results that directly report alien detection rates. Thus, there are significant theoretical and empirical gaps in our understanding of open category detection. In this paper, we take a step toward addressing this gap by studying a simple, but practically-relevant variant of open category detection. In our setting, we are provided with a "clean" training set that contains only the target categories of interest and an unlabeled "contaminated" training set that contains a fraction alpha of alien examples. Under the assumption that we know an upper bound on alpha we develop an algorithm with PAC-style guarantees on the alien detection rate, while aiming to minimize false alarms. Empirical results on synthetic and standard benchmark datasets demonstrate the regimes in which the algorithm can be effective and provide a baseline for further advancements.
****
Fast Variance Reduction Method with Stochastic Batch Size
Xuanqing Liu, Cho-Jui Hsieh
In this paper we study a family of variance reduction methods with randomized batch size—at each step, the algorithm first randomly chooses the batch size and then selects a batch of samples to conduct a variance-reduced stochastic update. We give the linear converge rate for this framework for composite functions, and show that the optimal strategy to achieve the best converge rate per data access is to always choose batch size equalling to 1, which is equivalent to the SAGA algorithm. However, due to the presence of cache/disk IO effect in computer architecture, number of data access cannot reflect the running time because of 1) random memory access is much slower than sequential access, 2) when data is too big to fit into memory, disk seeking takes even longer time. After taking these into account, choosing batch size equals to 1 is no longer optimal, so we propose a new algorithm called SAGA++ and theoretically show how to calculate the optimal average batch size. Our algorithm outperforms SAGA and other existing batch and stochastic solvers on real datasets. In addition, we also conduct a precise analysis to compare different update rules for variance reduction methods, showing that SAGA++ converges faster than SVRG in theory.
****
Fast Stochastic AUC Maximization with $O(1/n)$-Convergence Rate
Mingrui Liu, Xiaoxuan Zhang, Zaiyi Chen, Xiaoyu Wang, Tianbao Yang

In this paper, we consider statistical learning with AUC (area under ROC curve) maximization in the classical stochastic setting where one random data drawn fro m an unknown distribution is revealed at each iteration for updating the model. Although consistent convex surrogate losses for AUC maximization have been propo sed to make the problem tractable, it remains an challenging problem to design f ast optimization algorithms in the classical stochastic setting due to that the convex surrogate loss depends on random pairs of examples from positive and nega tive classes. Building on a saddle point formulation for a consistent square los s, this paper proposes a novel stochastic algorithm to improve the standard $O(1 /\sqrt{n})$ convergence rate to $\widetilde O(1/n)$ convergence rate without str ong convexity assumption or any favorable statistical assumptions (e.g., low noi se), where n is the number of random samples. To the best of our knowledge, th is is the first stochastic algorithm for AUC maximization with a statistical con vergence rate as fast as $O(1/n)$ up to a logarithmic factor. Extensive experime nts on eight large-scale benchmark data sets demonstrate the superior performanc e of the proposed algorithm comparing with existing stochastic or online algorit hms for AUC maximization.

****

On Matching Pursuit and Coordinate Descent

Francesco Locatello, Anant Raj, Sai Praneeth Karimireddy, Gunnar Raetsch, Bernha rd Schölkopf, Sebastian Stich, Martin Jaggi

Two popular examples of first-order optimization methods over linear spaces are coordinate descent and matching pursuit algorithms, with their randomized varian ts. While the former targets the optimization by moving along coordinates, the l atter considers a generalized notion of directions. Exploiting the connection be tween the two algorithms, we present a unified analysis of both, providing affin e invariant sublinear $O(1/t)$ rates on smooth objectives and linear convergence on strongly convex objectives. As a byproduct of our affine invariant analysis of matching pursuit, our rates for steepest coordinate descent are the tightest known. Furthermore, we show the first accelerated convergence rate $O(1/t^2)$ fo r matching pursuit and steepest coordinate descent on convex objectives.

****

PDE-Net: Learning PDEs from Data

Zichao Long, Yiping Lu, Xianzhong Ma, Bin Dong

Partial differential equations (PDEs) play a prominent role in many disciplines of science and engineering. PDEs are commonly derived based on empirical observa tions. However, with the rapid development of sensors, computational power, and data storage in the past decade, huge quantities of data can be easily collected and efficiently stored. Such vast quantity of data offers new opportunities for data-driven discovery of physical laws. Inspired by the latest development of n eural network designs in deep learning, we propose a new feed-forward deep netwo rk, called PDE-Net, to fulfill two objectives at the same time: to accurately pr edict dynamics of complex systems and to uncover the underlying hidden PDE model s. Comparing with existing approaches, our approach has the most flexibility by learning both differential operators and the nonlinear response function of the underlying PDE model. A special feature of the proposed PDE-Net is that all filt ers are properly constrained, which enables us to easily identify the governing PDE models while still maintaining the expressive and predictive power of the ne twork. These constrains are carefully designed by fully exploiting the relation between the orders of differential operators and the orders of sum rules of filt ers (an important concept originated from wavelet theory). Numerical experiments show that the PDE-Net has the potential to uncover the hidden PDE of the observ ed dynamics, and predict the dynamical behavior for a relatively long time, even in a noisy environment.

****

Error Estimation for Randomized Least-Squares Algorithms via the Bootstrap

Miles Lopes, Shusen Wang, Michael Mahoney

Over the course of the past decade, a variety of randomized algorithms have been proposed for computing approximate least-squares (LS) solutions in large-scale settings. A longstanding practical issue is that, for any given input, the user

rarely knows the actual error of an approximate solution (relative to the exact solution). Likewise, it is difficult for the user to know precisely how much computation is needed to achieve the desired error tolerance. Consequently, the user often appeals to worst-case error bounds that tend to offer only qualitative guidance. As a more practical alternative, we propose a bootstrap method to compute a posteriori error estimates for randomized LS algorithms. These estimates permit the user to numerically assess the error of a given solution, and to predict how much work is needed to improve a "preliminary" solution. In addition, we provide theoretical consistency results for the method, which are the first such results in this context (to the best of our knowledge). From a practical standpoint, the method also has considerable flexibility, insofar as it can be applied to several popular sketching algorithms, as well as a variety of error metrics. Moreover, the extra step of error estimation does not add much cost to an underlying sketching algorithm. Finally, we demonstrate the effectiveness of the method with empirical results.

****

Constraining the Dynamics of Deep Probabilistic Models

Marco Lorenzi, Maurizio Filippone

We introduce a novel generative formulation of deep probabilistic models implementing "soft" constraints on their function dynamics. In particular, we develop a flexible methodological framework where the modeled functions and derivatives of a given order are subject to inequality or equality constraints. We then characterize the posterior distribution over model and constraint parameters through stochastic variational inference. As a result, the proposed approach allows for accurate and scalable uncertainty quantification on the predictions and on all parameters. We demonstrate the application of equality constraints in the challenging problem of parameter inference in ordinary differential equation models, while we showcase the application of inequality constraints on the problem of monotonic regression of count data. The proposed approach is extensively tested in several experimental settings, leading to highly competitive results in challenging modeling applications, while offering high expressiveness, flexibility and scalability.

****

Spectrally Approximating Large Graphs with Smaller Graphs

Andreas Loukas, Pierre Vandergheynst

How does coarsening affect the spectrum of a general graph? We provide conditions such that the principal eigenvalues and eigenspaces of a coarsened and original graph Laplacian matrices are close. The achieved approximation is shown to depend on standard graph-theoretic properties, such as the degree and eigenvalue distributions, as well as on the ratio between the coarsened and actual graph sizes. Our results carry implications for learning methods that utilize coarsening. For the particular case of spectral clustering, they imply that coarse eigenvectors can be used to derive good quality assignments even without refinement{—}this phenomenon was previously observed, but lacked formal justification.

****

The Edge Density Barrier: Computational-Statistical Tradeoffs in Combinatorial Inference

Hao Lu, Yuan Cao, Zhuoran Yang, Junwei Lu, Han Liu, Zhaoran Wang

We study the hypothesis testing problem of inferring the existence of combinatorial structures in undirected graphical models. Although there exist extensive studies on the information-theoretic limits of this problem, it remains largely unexplored whether such limits can be attained by efficient algorithms. In this paper, we quantify the minimum computational complexity required to attain the information-theoretic limits based on an oracle computational model. We prove that, for testing common combinatorial structures, such as clique, nearest neighbor graph and perfect matching, against an empty graph, or large clique against small clique, the information-theoretic limits are provably unachievable by tractable algorithms in general. More importantly, we define structural quantities called the weak and strong edge densities, which offer deep insight into the existence of such computational-statistical tradeoffs. To the best of our knowledge, our

characterization is the first to identify and explain the fundamental tradeoffs between statistics and computation for combinatorial inference problems in undirected graphical models.

****

Accelerating Greedy Coordinate Descent Methods

Haihao Lu, Robert Freund, Vahab Mirrokni

We introduce and study two algorithms to accelerate greedy coordinate descent in theory and in practice: Accelerated Semi-Greedy Coordinate Descent (ASCD) and Accelerated Greedy Coordinate Descent (AGCD). On the theory side, our main results are for ASCD: we show that ASCD achieves $O(1/k^2)$ convergence, and it also achieves accelerated linear convergence for strongly convex functions. On the empirical side, while both AGCD and ASCD outperform Accelerated Randomized Coordinate Descent on most instances in our numerical experiments, we note that AGCD significantly outperforms the other two methods in our experiments, in spite of a lack of theoretical guarantees for this method. To complement this empirical finding for AGCD, we present an explanation why standard proof techniques for acceleration cannot work for AGCD, and we further introduce a technical condition under which AGCD is guaranteed to have accelerated convergence. Finally, we confirm that this technical condition holds in our numerical experiments.

****

Structured Variationally Auto-encoded Optimization

Xiaoyu Lu, Javier Gonzalez, Zhenwen Dai, Neil D. Lawrence

We tackle the problem of optimizing a black-box objective function defined over a highly-structured input space. This problem is ubiquitous in science and engineering. In machine learning, inferring the structure of a neural network or the Automatic Statistician (AS), where the optimal kernel combination for a Gaussian process is selected, are two important examples. We use the \as as a case study to describe our approach, that can be easily generalized to other domains. We propose an Structure Generating Variational Auto-encoder (SG-VAE) to embed the original space of kernel combinations into some low-dimensional continuous manifold where Bayesian optimization (BO) ideas are used. This is possible when structural knowledge of the problem is available, which can be given via a simulator or any other form of generating potentially good solutions. The right exploration-exploitation balance is imposed by propagating into the search the uncertainty of the latent space of the SG-VAE, that is computed using variational inference. The key aspect of our approach is that the SG-VAE can be used to bias the search towards relevant regions, making it suitable for transfer learning tasks. Several experiments in various application domains are used to illustrate the utility and generality of the approach described in this work.

****

Beyond Finite Layer Neural Networks: Bridging Deep Architectures and Numerical Differential Equations

Yiping Lu, Aoxiao Zhong, Quanzheng Li, Bin Dong

Deep neural networks have become the state-of-the-art models in numerous machine learning tasks. However, general guidance to network architecture design is still missing. In our work, we bridge deep neural network design with numerical differential equations. We show that many effective networks, such as ResNet, PolyNet, FractalNet and RevNet, can be interpreted as different numerical discretizations of differential equations. This finding brings us a brand new perspective on the design of effective deep architectures. We can take advantage of the rich knowledge in numerical analysis to guide us in designing new and potentially more effective deep networks. As an example, we propose a linear multi-step architecture (LM-architecture) which is inspired by the linear multi-step method solving ordinary differential equations. The LM-architecture is an effective structure that can be used on any ResNet-like networks. In particular, we demonstrate that LM-ResNet and LM-ResNeXt (i.e. the networks obtained by applying the LM-architecture on ResNet and ResNeXt respectively) can achieve noticeably higher accuracy than ResNet and ResNeXt on both CIFAR and ImageNet with comparable numbers of trainable parameters. In particular, on both CIFAR and ImageNet, LM-ResNet/LM-ResNeXt can significantly compress (>50%) the original networks while maintaining

a similar performance. This can be explained mathematically using the concept of modified equation from numerical analysis. Last but not least, we also establish a connection between stochastic control and noise injection in the training process which helps to improve generalization of the networks. Furthermore, by relating stochastic training strategy with stochastic dynamic system, we can easily apply stochastic training to the networks with the LM-architecture. As an example, we introduced stochastic depth to LM-ResNet and achieve significant improvement over the original LM-ResNet on CIFAR10.
****

End-to-end Active Object Tracking via Reinforcement Learning
Wenhan Luo, Peng Sun, Fangwei Zhong, Wei Liu, Tong Zhang, Yizhou Wang
We study active object tracking, where a tracker takes as input the visual observation (i.e. frame sequence) and produces the camera control signal (e.g., move forward, turn left, etc). Conventional methods tackle the tracking and the camera control separately, which is challenging to tune jointly. It also incurs many human efforts for labeling and many expensive trial-and-errors in real-world. To address these issues, we propose, in this paper, an end-to-end solution via deep reinforcement learning, where a ConvNet-LSTM function approximator is adopted for the direct frame-to-action prediction. We further propose an environment augmentation technique and a customized reward function, which are crucial for a successful training. The tracker trained in simulators (ViZDoom, Unreal Engine) shows good generalization in the case of unseen object moving path, unseen object appearance, unseen background, and distracting object. It can restore tracking when occasionally losing the target. With the experiments over the VOT dataset, we also find that the tracking ability, obtained solely from simulators, can potentially transfer to real-world scenarios.
****

Competitive Caching with Machine Learned Advice
Thodoris Lykouris, Sergei Vassilvtiskii
We develop a framework for augmenting online algorithms with a machine learned oracle to achieve competitive ratios that provably improve upon unconditional worst case lower bounds when the oracle has low error. Our approach treats the oracle as a complete black box, and is not dependent on its inner workings, or the exact distribution of its errors. We apply this framework to the traditional caching problem {—} creating an eviction strategy for a cache of size k. We demonstrate that naively following the oracle’s recommendations may lead to very poor performance, even when the average error is quite low. Instead we show how to modify the Marker algorithm to take into account the oracle’s predictions, and prove that this combined approach achieves a competitive ratio that both (i) decreases as the oracle’s error decreases, and (ii) is always capped by O(log k), which can be achieved without any oracle input. We complement our results with an empirical evaluation of our algorithm on real world datasets, and show that it performs well empirically even using simple off the shelf predictions.
****

Batch Bayesian Optimization via Multi-objective Acquisition Ensemble for Automated Analog Circuit Design
Wenlong Lyu, Fan Yang, Changhao Yan, Dian Zhou, Xuan Zeng
Bayesian optimization methods are promising for the optimization of black-box functions that are expensive to evaluate. In this paper, a novel batch Bayesian optimization approach is proposed. The parallelization is realized via a multi-objective ensemble of multiple acquisition functions. In each iteration, the multi-objective optimization of the multiple acquisition functions is performed to search for the Pareto front of the acquisition functions. The batch of inputs are then selected from the Pareto front. The Pareto front represents the best trade-off between the multiple acquisition functions. Such a policy for batch Bayesian optimization can significantly improve the efficiency of optimization. The proposed method is compared with several state-of-the-art batch Bayesian optimization algorithms using analytical benchmark functions and real-world analog integrated circuits. The experimental results show that the proposed method is competitive compared with the state-of-the-art algorithms.

****
Celer: a Fast Solver for the Lasso with Dual Extrapolation
Mathurin MASSIAS, Alexandre Gramfort, Joseph Salmon
Convex sparsity-inducing regularizations are ubiquitous in high-dimensional machine learning, but solving the resulting optimization problems can be slow. To accelerate solvers, state-of-the-art approaches consist in reducing the size of the optimization problem at hand. In the context of regression, this can be achieved either by discarding irrelevant features (screening techniques) or by prioritizing features likely to be included in the support of the solution (working set techniques). Duality comes into play at several steps in these techniques. Here, we propose an extrapolation technique starting from a sequence of iterates in the dual that leads to the construction of improved dual points. This enables a tighter control of optimality as used in stopping criterion, as well as better screening performance of Gap Safe rules. Finally, we propose a working set strategy based on an aggressive use of Gap Safe screening rules. Thanks to our new dual point construction, we show significant computational speedups on multiple real-world problems.
****
The Power of Interpolation: Understanding the Effectiveness of SGD in Modern Over-parametrized Learning
Siyuan Ma, Raef Bassily, Mikhail Belkin
In this paper we aim to formally explain the phenomenon of fast convergence of Stochastic Gradient Descent (SGD) observed in modern machine learning. The key observation is that most modern learning architectures are over-parametrized and are trained to interpolate the data by driving the empirical loss (classification and regression) close to zero. While it is still unclear why these interpolated solutions perform well on test data, we show that these regimes allow for fast convergence of SGD, comparable in number of iterations to full gradient descent. For convex loss functions we obtain an exponential convergence bound for mini-batch SGD parallel to that for full gradient descent. We show that there is a critical batch size m^* such that: (a) SGD iteration with mini-batch size $m\leq m^*$ is nearly equivalent to m iterations of mini-batch size 1 (linear scaling regime). (b) SGD iteration with mini-batch $m> m^*$ is nearly equivalent to a full gradient descent iteration (saturation regime). Moreover, for the quadratic loss, we derive explicit expressions for the optimal mini-batch and step size and explicitly characterize the two regimes above. The critical mini-batch size can be viewed as the limit for effective mini-batch parallelization. It is also nearly independent of the data size, implying $O(n)$ acceleration over GD per unit of computation. We give experimental evidence on real data which closely follows our theoretical analyses. Finally, we show how our results fit in the recent developments in training deep neural networks and discuss connections to adaptive rates for SGD and variance reduction.
****
Gradient Descent for Sparse Rank-One Matrix Completion for Crowd-Sourced Aggregation of Sparsely Interacting Workers
Yao Ma, Alexander Olshevsky, Csaba Szepesvari, Venkatesh Saligrama
We consider worker skill estimation for the single coin Dawid-Skene crowdsourcing model. In practice skill-estimation is challenging because worker assignments are sparse and irregular due to the arbitrary, and uncontrolled availability of workers. We formulate skill estimation as a rank-one correlation-matrix completion problem, where the observed components correspond to observed label correlation between workers. We show that the correlation matrix can be successfully recovered and skills identifiable if and only if the sampling matrix (observed components) is irreducible and aperiodic. We then propose an efficient gradient descent scheme and show that skill estimates converges to the desired global optima for such sampling matrices. Our proof is original and the results are surprising in light of the fact that even the weighted rank-one matrix factorization problem is NP hard in general. Next we derive sample complexity bounds for the noisy case in terms of spectral properties of the signless Laplacian of the sampling matrix. Our proposed scheme achieves state-of-art performance on a number of real-

world datasets.
****

Implicit Regularization in Nonconvex Statistical Estimation: Gradient Descent Co nverges Linearly for Phase Retrieval and Matrix Completion
Cong Ma, Kaizheng Wang, Yuejie Chi, Yuxin Chen
Recent years have seen a flurry of activities in designing provably efficient no nconvex optimization procedures for solving statistical estimation problems. For various problems like phase retrieval or low-rank matrix completion, state-of-t he-art nonconvex procedures require proper regularization (e.g. trimming, regula rized cost, projection) in order to guarantee fast convergence. When it comes to vanilla procedures such as gradient descent, however, prior theory either recom mends highly conservative learning rates to avoid overshooting, or completely la cks performance guarantees. This paper uncovers a striking phenomenon in several nonconvex problems: even in the absence of explicit regularization, gradient de scent follows a trajectory staying within a basin that enjoys nice geometry, con sisting of points incoherent with the sampling mechanism. This “implicit regular ization” feature allows gradient descent to proceed in a far more aggressive fas hion without overshooting, which in turn results in substantial computational sa vings. Focusing on two statistical estimation problems, i.e. solving random quad ratic systems of equations and low-rank matrix completion, we establish that gra dient descent achieves near-optimal statistical and computational guarantees wit hout explicit regularization. As a byproduct, for noisy matrix completion, we de monstrate that gradient descent enables optimal control of both entrywise and sp ectral-norm errors.
****

Dimensionality-Driven Learning with Noisy Labels
Xingjun Ma, Yisen Wang, Michael E. Houle, Shuo Zhou, Sarah Erfani, Shutao Xia, S udanthi Wijewickrema, James Bailey
Datasets with significant proportions of noisy (incorrect) class labels present challenges for training accurate Deep Neural Networks (DNNs). We propose a new p erspective for understanding DNN generalization for such datasets, by investigat ing the dimensionality of the deep representation subspace of training samples. We show that from a dimensionality perspective, DNNs exhibit quite distinctive l earning styles when trained with clean labels versus when trained with a proport ion of noisy labels. Based on this finding, we develop a new dimensionality-driv en learning strategy, which monitors the dimensionality of subspaces during trai ning and adapts the loss function accordingly. We empirically demonstrate that o ur approach is highly tolerant to significant proportions of noisy labels, and c an effectively learn low-dimensional local subspaces that capture the data distr ibution.
****

Approximate message passing for amplitude based optimization
Junjie Ma, Ji Xu, Arian Maleki
We consider an ℓ_2-regularized non-convex optimization problem for recoveri ng signals from their noisy phaseless observations. We design and study the perf ormance of a message passing algorithm that aims to solve this optimization prob lem. We consider the asymptotic setting $m,n \rightarrow \infty$, $m/n \rightarr ow \delta$ and obtain sharp performance bounds, where m is the number of measu rements and n is the signal dimension. We show that for complex signals the al gorithm can perform accurate recovery with only $m=\left (\frac{64}{\pi^2}-4\ri ght)n\approx 2.5n$ measurements. Also, we provide sharp analysis on the sensitiv ity of the algorithm to noise. We highlight the following facts about our messag e passing algorithm: (i) Adding ℓ_2 regularization to the non-convex loss f unction can be beneficial even in the noiseless setting; (ii) spectral initializ ation has marginal impact on the performance of the algorithm.
****

Orthogonal Machine Learning: Power and Limitations
Lester Mackey, Vasilis Syrgkanis, Ilias Zadik
Double machine learning provides n^{1/2}-consistent estimates of parameters of i nterest even when high-dimensional or nonparametric nuisance parameters are esti

mated at an n^{-1/4} rate. The key is to employ Neyman-orthogonal moment equatio ns which are first-order insensitive to perturbations in the nuisance parameters . We show that the n^{-1/4} requirement can be improved to n^{-1/(2k+2)} by empl oying a k-th order notion of orthogonality that grants robustness to more comple x or higher-dimensional nuisance parameters. In the partially linear regression setting popular in causal inference, we show that we can construct second-order orthogonal moments if and only if the treatment residual is not normally distrib uted. Our proof relies on Stein’s lemma and may be of independent interest. We c onclude by demonstrating the robustness benefits of an explicit doubly-orthogona l estimation procedure for treatment effect.
****

Learning Adversarially Fair and Transferable Representations
David Madras, Elliot Creager, Toniann Pitassi, Richard Zemel
In this paper, we advocate for representation learning as the key to mitigating unfair prediction outcomes downstream. Motivated by a scenario where learned rep resentations are used by third parties with unknown objectives, we propose and e xplore adversarial representation learning as a natural method of ensuring those parties act fairly. We connect group fairness (demographic parity, equalized od ds, and equal opportunity) to different adversarial objectives. Through worst-ca se theoretical guarantees and experimental validation, we show that the choice o f this objective is crucial to fair prediction. Furthermore, we present the firs t in-depth experimental demonstration of fair transfer learning and demonstrate empirically that our learned representations admit fair predictions on new tasks while maintaining utility, an essential goal of fair representation learning.
****

An Efficient, Generalized Bellman Update For Cooperative Inverse Reinforcement L earning
Dhruv Malik, Malayandi Palaniappan, Jaime Fisac, Dylan Hadfield-Menell, Stuart R ussell, Anca Dragan
Our goal is for AI systems to correctly identify and act according to their huma n user’s objectives. Cooperative Inverse Reinforcement Learning (CIRL) formalize s this value alignment problem as a two-player game between a human and robot, i n which only the human knows the parameters of the reward function: the robot ne eds to learn them as the interaction unfolds. Previous work showed that CIRL can be solved as a POMDP, but with an action space size exponential in the size of the reward parameter space. In this work, we exploit a specific property of CIRL : the human is a full information agent. This enables us to derive an optimality -preserving modification to the standard Bellman update, which reduces the compl exity of the problem by an exponential factor. Additionally, we show that our mo dified Bellman update allows us to relax CIRL’s assumption of human rationality. We apply this update to a variety of POMDP solvers, including exact methods, po int-based methods, and Monte Carlo Tree Search methods. We find that it enables us to scale CIRL to non-trivial problems, with larger reward parameter spaces, a nd larger action spaces for both robot and human. In solutions to these larger p roblems, the human exhibits pedagogical (teaching) behavior, while the robot int erprets it as such and attains higher value for the human.
****

Iterative Amortized Inference
Joe Marino, Yisong Yue, Stephan Mandt
Inference models are a key component in scaling variational inference to deep la tent variable models, most notably as encoder networks in variational auto-encod ers (VAEs). By replacing conventional optimization-based inference with a learne d model, inference is amortized over data examples and therefore more computatio nally efficient. However, standard inference models are restricted to direct map pings from data to approximate posterior estimates. The failure of these models to reach fully optimized approximate posterior estimates results in an amortizat ion gap. We aim toward closing this gap by proposing iterative inference models, which learn to perform inference optimization through repeatedly encoding gradi ents. Our approach generalizes standard inference models in VAEs and provides in sight into several empirical findings, including top-down inference techniques.

We demonstrate the inference optimization capabilities of iterative inference mo
dels and show that they outperform standard inference models on several benchmar
k data sets of images and text.
****

Streaming Principal Component Analysis in Noisy Setting
Teodor Vanislavov Marinov, Poorya Mianjy, Raman Arora
We study streaming algorithms for principal component analysis (PCA) in noisy se
ttings. We present computationally efficient algorithms with sub-linear regret b
ounds for PCA in the presence of noise, missing data, and gross outliers.
****

Fast Approximate Spectral Clustering for Dynamic Networks
Lionel Martin, Andreas Loukas, Pierre Vandergheynst
Spectral clustering is a widely studied problem, yet its complexity is prohibiti
ve for dynamic graphs of even modest size. We claim that it is possible to reuse
 information of past cluster assignments to expedite computation. Our approach b
uilds on a recent idea of sidestepping the main bottleneck of spectral clusterin
g, i.e., computing the graph eigenvectors, by a polynomial-based randomized sket
ching technique. We show that the proposed algorithm achieves clustering assignm
ents with quality approximating that of spectral clustering and that it can yiel
d significant complexity benefits when the graph dynamics are appropriately boun
ded. In our experiments, our method clusters 30k node graphs 3.9$\times$ faster
in average and deviates from the correct assignment by less than 0.1%.
****

Bayesian Model Selection for Change Point Detection and Clustering
Othmane Mazhar, Cristian Rojas, Carlo Fischione, Mohammad Reza Hesamzadeh
We address a generalization of change point detection with the purpose of detect
ing the change locations and the levels of clusters of a piecewise constant sign
al. Our approach is to model it as a nonparametric penalized least square model
selection on a family of models indexed over the collection of partitions of the
 design points and propose a computationally efficient algorithm to approximatel
y solve it. Statistically, minimizing such a penalized criterion yields an appro
ximation to the maximum a-posteriori probability (MAP) estimator. The criterion
is then analyzed and an oracle inequality is derived using a Gaussian concentrat
ion inequality. The oracle inequality is used to derive on one hand conditions f
or consistency and on the other hand an adaptive upper bound on the expected squ
are risk of the estimator, which statistically motivates our approximation. Fina
lly, we apply our algorithm to simulated data to experimentally validate the sta
tistical guarantees and illustrate its behavior.
****

Optimization, fast and slow: optimally switching between local and Bayesian opti
mization
Mark McLeod, Stephen Roberts, Michael A. Osborne
We develop the first Bayesian Optimization algorithm, BLOSSOM, which selects bet
ween multiple alternative acquisition functions and traditional local optimizati
on at each step. This is combined with a novel stopping condition based on expec
ted regret. This pairing allows us to obtain the best characteristics of both lo
cal and Bayesian optimization, making efficient use of function evaluations whil
e yielding superior convergence to the global minimum on a selection of optimiza
tion problems, and also halting optimization once a principled and intuitive sto
pping condition has been fulfilled.
****

Bounds on the Approximation Power of Feedforward Neural Networks
Mohammad Mehrabi, Aslan Tchamkerten, MANSOOR YOUSEFI
The approximation power of general feedforward neural networks with piecewise li
near activation functions is investigated. First, lower bounds on the size of a
network are established in terms of the approximation error and network depth an
d width. These bounds improve upon state-of-the-art bounds for certain classes o
f functions, such as strongly convex functions. Second, an upper bound is establ
ished on the difference of two neural networks with identical weights but differ
ent activation functions.

****
Differentiable Dynamic Programming for Structured Prediction and Attention
Arthur Mensch, Mathieu Blondel
Dynamic programming (DP) solves a variety of structured combinatorial problems by iteratively breaking them down into smaller subproblems. In spite of their versatility, many DP algorithms are non-differentiable, which hampers their use as a layer in neural networks trained by backpropagation. To address this issue, we propose to smooth the max operator in the dynamic programming recursion, using a strongly convex regularizer. This allows to relax both the optimal value and solution of the original combinatorial problem, and turns a broad class of DP algorithms into differentiable operators. Theoretically, we provide a new probabilistic perspective on backpropagating through these DP operators, and relate them to inference in graphical models. We derive two particular instantiations of our framework, a smoothed Viterbi algorithm for sequence prediction and a smoothed DTW algorithm for time-series alignment. We showcase these instantiations on structured prediction (audio-to-score alignment, NER) and on structured and sparse attention for translation.
****
Ranking Distributions based on Noisy Sorting
Adil El Mesaoudi-Paul, Eyke Hüllermeier, Robert Busa-Fekete
We propose a new statistical model for ranking data, i.e., a new family of probability distributions on permutations. Our model is inspired by the idea of a data-generating process in the form of a noisy sorting procedure, in which deterministic comparisons between pairs of items are replaced by Bernoulli trials. The probability of producing a certain ranking as a result then essentially depends on the Bernoulli parameters, which can be interpreted as pairwise preferences. We show that our model can be written in closed form if insertion or quick sort are used as sorting algorithms, and propose a maximum likelihood approach for parameter estimation. We also introduce a generalization of the model, in which the constraints on pairwise preferences are relaxed, and for which maximum likelihood estimation can be carried out based on a variation of the generalized iterative scaling algorithm. Experimentally, we show that the models perform very well in terms of goodness of fit, compared to existing models for ranking data.
****
Which Training Methods for GANs do actually Converge?
Lars Mescheder, Andreas Geiger, Sebastian Nowozin
Recent work has shown local convergence of GAN training for absolutely continuous data and generator distributions. In this paper, we show that the requirement of absolute continuity is necessary: we describe a simple yet prototypical counterexample showing that in the more realistic case of distributions that are not absolutely continuous, unregularized GAN training is not always convergent. Furthermore, we discuss regularization strategies that were recently proposed to stabilize GAN training. Our analysis shows that GAN training with instance noise or zero-centered gradient penalties converges. On the other hand, we show that Wasserstein-GANs and WGAN-GP with a finite number of discriminator updates per generator update do not always converge to the equilibrium point. We discuss these results, leading us to a new explanation for the stability problems of GAN training. Based on our analysis, we extend our convergence results to more general GANs and prove local convergence for simplified gradient penalties even if the generator and data distributions lie on lower dimensional manifolds. We find these penalties to work well in practice and use them to learn high-resolution generative image models for a variety of datasets with little hyperparameter tuning.
****
Configurable Markov Decision Processes
Alberto Maria Metelli, Mirco Mutti, Marcello Restelli
In many real-world problems, there is the possibility to configure, to a limited extent, some environmental parameters to improve the performance of a learning agent. In this paper, we propose a novel framework, Configurable Markov Decision Processes (Conf-MDPs), to model this new type of interaction with the environment. Furthermore, we provide a new learning algorithm, Safe Policy-Model Iteratio

n (SPMI), to jointly and adaptively optimize the policy and the environment conf iguration. After having introduced our approach and derived some theoretical res ults, we present the experimental evaluation in two explicative problems to show the benefits of the environment configurability on the performance of the learn ed policy.
****

prDeep: Robust Phase Retrieval with a Flexible Deep Network
Christopher Metzler, Phillip Schniter, Ashok Veeraraghavan, Richard Baraniuk
Phase retrieval algorithms have become an important component in many modern com putational imaging systems. For instance, in the context of ptychography and spe ckle correlation imaging, they enable imaging past the diffraction limit and thr ough scattering media, respectively. Unfortunately, traditional phase retrieval algorithms struggle in the presence of noise. Progress has been made recently on developing more robust algorithms using signal priors, but at the expense of li miting the range of supported measurement models (e.g., to Gaussian or coded dif fraction patterns). In this work we leverage the regularization-by-denoising fra mework and a convolutional neural network denoiser to create prDeep, a new phase retrieval algorithm that is both robust and broadly applicable. We test and val idate prDeep in simulation to demonstrate that it is robust to noise and can han dle a variety of system models.
****

Pseudo-task Augmentation: From Deep Multitask Learning to Intratask Sharing—and Back
Elliot Meyerson, Risto Miikkulainen
Deep multitask learning boosts performance by sharing learned structure across r elated tasks. This paper adapts ideas from deep multitask learning to the settin g where only a single task is available. The method is formalized as pseudo-task augmentation, in which models are trained with multiple decoders for each task. Pseudo-tasks simulate the effect of training towards closely-related tasks draw n from the same universe. In a suite of experiments, pseudo-task augmentation is shown to improve performance on single-task learning problems. When combined wi th multitask learning, further improvements are achieved, including state-of-the -art performance on the CelebA dataset, showing that pseudo-task augmentation an d multitask learning have complementary value. All in all, pseudo-task augmentat ion is a broadly applicable and efficient way to boost performance in deep learn ing systems.
****

The Hidden Vulnerability of Distributed Learning in Byzantium
El Mahdi El Mhamdi, Rachid Guerraoui, Sébastien Rouault
While machine learning is going through an era of celebrated success, concerns h ave been raised about the vulnerability of its backbone: stochastic gradient des cent (SGD). Recent approaches have been proposed to ensure the robustness of dis tributed SGD against adversarial (Byzantine) workers sending poisoned gradients during the training phase. Some of these approaches have been proven Byzantine–r esilient: they ensure the convergence of SGD despite the presence of a minority of adversarial workers. We show in this paper that convergence is not enough. In high dimension $d \gg 1$, an adver\-sary can build on the loss function’s non–c onvexity to make SGD converge to ineffective models. More precisely, we bring to light that existing Byzantine–resilient schemes leave a margin of poisoning of $\bigOmega\left(f(d)\right)$, where $f(d)$ increases at least like $\sqrt[p]{d } $. Based on this leeway, we build a simple attack, and experimentally show its s trong to utmost effectivity on CIFAR–10 and MNIST. We introduce Bulyan, and prov e it significantly reduces the attackers leeway to a narrow $\bigO\,(\sfrac{1}{ \sqrt{d }})$ bound. We empirically show that Bulyan does not suffer the fragilit y of existing aggregation rules and, at a reasonable cost in terms of required b atch size, achieves convergence as if only non–Byzantine gradients had been used to update the model.
****

Stochastic PCA with ℓ_2 and ℓ_1 Regularization
Poorya Mianjy, Raman Arora

We revisit convex relaxation based methods for stochastic optimization of principal component analysis (PCA). While methods that directly solve the nonconvex problem have been shown to be optimal in terms of statistical and computational efficiency, the methods based on convex relaxation have been shown to enjoy comparable, or even superior, empirical performance – this motivates the need for a deeper formal understanding of the latter. Therefore, in this paper, we study variants of stochastic gradient descent for a convex relaxation of PCA with (a) ℓ_2, (b) ℓ_1, and (c) elastic net ($\ell_1+\ell_2)$ regularization in the hope that these variants yield (a) better iteration complexity, (b) better control on the rank of the intermediate iterates, and (c) both, respectively. We show, theoretically and empirically, that compared to previous work on convex relaxation based methods, the proposed variants yield faster convergence and improve overall runtime to achieve a certain user-specified ϵ-suboptimality on the PCA objective. Furthermore, the proposed methods are shown to converge both in terms of the PCA objective as well as the distance between subspaces. However, there still remains a gap in computational requirements for the proposed methods when compared with existing nonconvex approaches.
****

On the Implicit Bias of Dropout
Poorya Mianjy, Raman Arora, Rene Vidal
Algorithmic approaches endow deep learning systems with implicit bias that helps them generalize even in over-parametrized settings. In this paper, we focus on understanding such a bias induced in learning through dropout, a popular technique to avoid overfitting in deep learning. For single hidden-layer linear neural networks, we show that dropout tends to make the norm of incoming/outgoing weight vectors of all the hidden nodes equal. In addition, we provide a complete characterization of the optimization landscape induced by dropout.
****

One-Shot Segmentation in Clutter
Claudio Michaelis, Matthias Bethge, Alexander Ecker
We tackle the problem of one-shot segmentation: finding and segmenting a previously unseen object in a cluttered scene based on a single instruction example. We propose a novel dataset, which we call cluttered Omniglot. Using a baseline architecture combining a Siamese embedding for detection with a U-net for segmentation we show that increasing levels of clutter make the task progressively harder. Using oracle models with access to various amounts of ground-truth information, we evaluate different aspects of the problem and show that in this kind of visual search task, detection and segmentation are two intertwined problems, the solution to each of which helps solving the other. We therefore introduce MaskNet, an improved model that attends to multiple candidate locations, generates segmentation proposals to mask out background clutter and selects among the segmented objects. Our findings suggest that such image recognition models based on an iterative refinement of object detection and foreground segmentation may provide a way to deal with highly cluttered scenes.
****

Differentiable plasticity: training plastic neural networks with backpropagation
Thomas Miconi, Kenneth Stanley, Jeff Clune
How can we build agents that keep learning from experience, quickly and efficiently, after their initial training? Here we take inspiration from the main mechanism of learning in biological brains: synaptic plasticity, carefully tuned by evolution to produce efficient lifelong learning. We show that plasticity, just like connection weights, can be optimized by gradient descent in large (millions of parameters) recurrent networks with Hebbian plastic connections. First, recurrent plastic networks with more than two million parameters can be trained to memorize and reconstruct sets of novel, high-dimensional (1000+ pixels) natural images not seen during training. Crucially, traditional non-plastic recurrent networks fail to solve this task. Furthermore, trained plastic networks can also solve generic meta-learning tasks such as the Omniglot task, with competitive results and little parameter overhead. Finally, in reinforcement learning settings, plastic networks outperform non-plastic equivalent in a maze exploration task. We

conclude that differentiable plasticity may provide a powerful novel approach to the learning-to-learn problem.
****

Training Neural Machines with Trace-Based Supervision
Matthew Mirman, Dimitar Dimitrov, Pavle Djordjevic, Timon Gehr, Martin Vechev
We investigate the effectiveness of trace-based supervision methods for training existing neural abstract machines. To define the class of neural machines amenable to trace-based supervision, we introduce the concept of a differential neural computational machine (dNCM) and show that several existing architectures (NTMs, NRAMs) can be described as dNCMs. We performed a detailed experimental evaluation with NTM and NRAM machines, showing that additional supervision on the interpretable portions of these architectures leads to better convergence and generalization capabilities of the learning phase than standard training, in both noise-free and noisy scenarios.
****

Differentiable Abstract Interpretation for Provably Robust Neural Networks
Matthew Mirman, Timon Gehr, Martin Vechev
We introduce a scalable method for training robust neural networks based on abstract interpretation. We present several abstract transformers which balance efficiency with precision and show these can be used to train large neural networks that are certifiably robust to adversarial perturbations.
****

A Delay-tolerant Proximal-Gradient Algorithm for Distributed Learning
Konstantin Mishchenko, Franck Iutzeler, Jérôme Malick, Massih-Reza Amini
Distributed learning aims at computing high-quality models by training over scattered data. This covers a diversity of scenarios, including computer clusters or mobile agents. One of the main challenges is then to deal with heterogeneous machines and unreliable communications. In this setting, we propose and analyze a flexible asynchronous optimization algorithm for solving nonsmooth learning problems. Unlike most existing methods, our algorithm is adjustable to various levels of communication costs, machines computational powers, and data distribution evenness. We prove that the algorithm converges linearly with a fixed learning rate that does not depend on communication delays nor on the number of machines. Although long delays in communication may slow down performance, no delay can break convergence.
****

Data Summarization at Scale: A Two-Stage Submodular Approach
Marko Mitrovic, Ehsan Kazemi, Morteza Zadimoghaddam, Amin Karbasi
The sheer scale of modern datasets has resulted in a dire need for summarization techniques that can identify representative elements in a dataset. Fortunately, the vast majority of data summarization tasks satisfy an intuitive diminishing returns condition known as submodularity, which allows us to find nearly-optimal solutions in linear time. We focus on a two-stage submodular framework where the goal is to use some given training functions to reduce the ground set so that optimizing new functions (drawn from the same distribution) over the reduced set provides almost as much value as optimizing them over the entire ground set. In this paper, we develop the first streaming and distributed solutions to this problem. In addition to providing strong theoretical guarantees, we demonstrate both the utility and efficiency of our algorithms on real-world tasks including image summarization and ride-share optimization.
****

The Hierarchical Adaptive Forgetting Variational Filter
Vincent Moens
A common problem in Machine Learning and statistics consists in detecting whether the current sample in a stream of data belongs to the same distribution as previous ones, is an isolated outlier or inaugurates a new distribution of data. We present a hierarchical Bayesian algorithm that aims at learning a time-specific approximate posterior distribution of the parameters describing the distribution of the data observed. We derive the update equations of the variational parameters of the approximate posterior at each time step for models from the exponent

ial family, and show that these updates find interesting correspondents in Reinf orcement Learning (RL). In this perspective, our model can be seen as a hierarch ical RL algorithm that learns a posterior distribution according to a certain st ability confidence that is, in turn, learned according to its own stability conf idence. Finally, we show some applications of our generic model, first in a RL c ontext, next with an adaptive Bayesian Autoregressive model, and finally in the context of Stochastic Gradient Descent optimization.
****

Decentralized Submodular Maximization: Bridging Discrete and Continuous Settings
Aryan Mokhtari, Hamed Hassani, Amin Karbasi
In this paper, we showcase the interplay between discrete and continuous optimiz ation in network-structured settings. We propose the first fully decentralized o ptimization method for a wide class of non-convex objective functions that posse ss a diminishing returns property. More specifically, given an arbitrary connect ed network and a global continuous submodular function, formed by a sum of local functions, we develop Decentralized Continuous Greedy (DCG), a message passing algorithm that converges to the tight $(1-1/e)$ approximation factor of the opti mum global solution using only local computation and communication. We also prov ide strong convergence bounds as a function of network size and spectral charact eristics of the underlying topology. Interestingly, DCG readily provides a simpl e recipe for decentralized discrete submodular maximization through the means of continuous relaxations. Formally, we demonstrate that by lifting the local disc rete functions to continuous domains and using DCG as an interface we can develo p a consensus algorithm that also achieves the tight $(1-1/e)$ approximation gua rantee of the global discrete solution once a proper rounding scheme is applied.
****

DICOD: Distributed Convolutional Coordinate Descent for Convolutional Sparse Cod ing
Thomas Moreau, Laurent Oudre, Nicolas Vayatis
In this paper, we introduce DICOD, a convolutional sparse coding algorithm which builds shift invariant representations for long signals. This algorithm is desi gned to run in a distributed setting, with local message passing, making it comm unication efficient. It is based on coordinate descent and uses locally greedy u pdates which accelerate the resolution compared to greedy coordinate selection. We prove the convergence of this algorithm and highlight its computational speed -up which is super-linear in the number of cores used. We also provide empirical evidence for the acceleration properties of our algorithm compared to state-of- the-art methods.
****

WHInter: A Working set algorithm for High-dimensional sparse second order Intera ction models
Marine Le Morvan, Jean-Philippe Vert
Learning sparse linear models with two-way interactions is desirable in many app lication domains such as genomics. ℓ_1-regularised linear models are popula r to estimate sparse models, yet standard implementations fail to address specif ically the quadratic explosion of candidate two-way interactions in high dimensi ons, and typically do not scale to genetic data with hundreds of thousands of fe atures. Here we present WHInter, a working set algorithm to solve large ℓ_1 -regularised problems with two-way interactions for binary design matrices. The novelty of WHInter stems from a new bound to efficiently identify working sets w hile avoiding to scan all features, and on fast computations inspired from solut ions to the maximum inner product search problem. We apply WHInter to simulated and real genetic data and show that it is more scalable and two orders of magnit ude faster than the state of the art.
****

Dropout Training, Data-dependent Regularization, and Generalization Bounds
Wenlong Mou, Yuchen Zhou, Jun Gao, Liwei Wang
We study the problem of generalization guarantees for dropout training. A genera l framework is first proposed for learning procedures with random perturbation o n model parameters. The generalization error is bounded by sum of two offset Rad

emacher complexities: the main term is Rademacher complexity of the hypothesis c lass with minus offset induced by the perturbation variance, which characterizes data-dependent regularization by the random perturbation; the auxiliary term is offset Rademacher complexity for the variance class, controlling the degree to which this regularization effect can be weakened. For neural networks, we estima te upper and lower bounds for the variance induced by truthful dropout, a varian t of dropout that we propose to ensure unbiased output and fit into our framewor k, and the variance bounds exhibits connection to adaptive regularization method s. By applying our framework to ReLU networks with one hidden layer, a generaliz ation upper bound is derived with no assumptions on the parameter norms or data distribution, with $O(1/n)$ fast rate and adaptivity to geometry of data points being achieved at the same time.

****

Kernelized Synaptic Weight Matrices

Lorenz Muller, Julien Martel, Giacomo Indiveri

In this paper we introduce a novel neural network architecture, in which weight matrices are re-parametrized in terms of low-dimensional vectors, interacting th rough kernel functions. A layer of our network can be interpreted as introducing a (potentially infinitely wide) linear layer between input and output. We descr ibe the theory underpinning this model and validate it with concrete examples, e xploring how it can be used to impose structure on neural networks in diverse ap plications ranging from data visualization to recommender systems. We achieve st ate-of-the-art performance in a collaborative filtering task (MovieLens).

****

Rapid Adaptation with Conditionally Shifted Neurons

Tsendsuren Munkhdalai, Xingdi Yuan, Soroush Mehri, Adam Trischler

We describe a mechanism by which artificial neural networks can learn rapid adap tation - the ability to adapt on the fly, with little data, to new tasks - that we call conditionally shifted neurons. We apply this mechanism in the framework of metalearning, where the aim is to replicate some of the flexibility of human learning in machines. Conditionally shifted neurons modify their activation valu es with task-specific shifts retrieved from a memory module, which is populated rapidly based on limited task experience. On metalearning benchmarks from the vi sion and language domains, models augmented with conditionally shifted neurons a chieve state-of-the-art results.

****

On the Relationship between Data Efficiency and Error for Uncertainty Sampling

Stephen Mussmann, Percy Liang

While active learning offers potential cost savings, the actual data efficiency—the reduction in amount of labeled data needed to obtain the same error rate—obs erved in practice is mixed. This paper poses a basic question: when is active le arning actually helpful? We provide an answer for logistic regression with the p opular active learning algorithm, uncertainty sampling. Empirically, on 21 datas ets from OpenML, we find a strong inverse correlation between data efficiency an d the error rate of the final classifier. Theoretically, we show that for a vari ant of uncertainty sampling, the asymptotic data efficiency is within a constant factor of the inverse error rate of the limiting classifier.

****

Fitting New Speakers Based on a Short Untranscribed Sample

Eliya Nachmani, Adam Polyak, Yaniv Taigman, Lior Wolf

Learning-based Text To Speech systems have the potential to generalize from one speaker to the next and thus require a relatively short sample of any new voice. However, this promise is currently largely unrealized. We present a method that is designed to capture a new speaker from a short untranscribed audio sample. T his is done by employing an additional network that given an audio sample, place s the speaker in the embedding space. This network is trained as part of the spe ech synthesis system using various consistency losses. Our results demonstrate a greatly improved performance on both the dataset speakers, and, more importantl y, when fitting new voices, even from very short samples.

****

Smoothed Action Value Functions for Learning Gaussian Policies
Ofir Nachum, Mohammad Norouzi, George Tucker, Dale Schuurmans
State-action value functions (i.e., Q-values) are ubiquitous in reinforcement le arning (RL), giving rise to popular algorithms such as SARSA and Q-learning. We propose a new notion of action value defined by a Gaussian smoothed version of t he expected Q-value. We show that such smoothed Q-values still satisfy a Bellman equation, making them learnable from experience sampled from an environment. Mo reover, the gradients of expected reward with respect to the mean and covariance of a parameterized Gaussian policy can be recovered from the gradient and Hessi an of the smoothed Q-value function. Based on these relationships we develop new algorithms for training a Gaussian policy directly from a learned smoothed Q-va lue approximator. The approach is additionally amenable to proximal optimization by augmenting the objective with a penalty on KL-divergence from a previous pol icy. We find that the ability to learn both a mean and covariance during trainin g leads to significantly improved results on standard continuous control benchma rks.
****
Nearly Optimal Robust Subspace Tracking
Praneeth Narayanamurthy, Namrata Vaswani
Robust subspace tracking (RST) can be simply understood as a dynamic (time-varyi ng) extension of robust PCA. More precisely, it is the problem of tracking data lying in a fixed or slowly-changing low-dimensional subspace while being robust to sparse outliers. This work develops a recursive projected compressive sensing algorithm called “Nearly Optimal RST (NORST)”, and obtains one of the first gua rantees for it. We show that NORST provably solves RST under weakened standard R PCA assumptions, slow subspace change, and a lower bound on (most) outlier magni tudes. Our guarantee shows that (i) NORST is online (after initialization) and e njoys near-optimal values of tracking delay, lower bound on required delay betwe en subspace change times, and of memory complexity; and (ii) it has a significan tly improved worst-case outlier tolerance compared with all previous robust PCA or RST methods without requiring any model on how the outlier support is generat ed.
****
Stochastic Proximal Algorithms for AUC Maximization
Michael Natole, Yiming Ying, Siwei Lyu
Stochastic optimization algorithms such as SGDs update the model sequentially wi th cheap per-iteration costs, making them amenable for large-scale data analysis . However, most of the existing studies focus on the classification accuracy whi ch can not be directly applied to the important problems of maximizing the Area under the ROC curve (AUC) in imbalanced classification and bipartite ranking. In this paper, we develop a novel stochastic proximal algorithm for AUC maximizati on which is referred to as SPAM. Compared with the previous literature, our algo rithm SPAM applies to a non-smooth penalty function, and achieves a convergence rate of O(log t/t) for strongly convex functions while both space and per-iterat ion costs are of one datum.
****
Mitigating Bias in Adaptive Data Gathering via Differential Privacy
Seth Neel, Aaron Roth
Data that is gathered adaptively — via bandit algorithms, for example — exhibits bias. This is true both when gathering simple numeric valued data — the empiric al means kept track of by stochastic bandit algorithms are biased downwards — an d when gathering more complicated data — running hypothesis tests on complex dat a gathered via contextual bandit algorithms leads to false discovery. In this pa per, we show that this problem is mitigated if the data collection procedure is differentially private. This lets us both bound the bias of simple numeric value d quantities (like the empirical means of stochastic bandit algorithms), and cor rect the p-values of hypothesis tests run on the adaptively gathered data. Moreo ver, there exist differentially private bandit algorithms with near optimal regr et bounds: we apply existing theorems in the simple stochastic case, and give a new analysis for linear contextual bandits. We complement our theoretical result

s with experiments validating our theory.
****
Optimization Landscape and Expressivity of Deep CNNs
Quynh Nguyen, Matthias Hein
We analyze the loss landscape and expressiveness of practical deep convolutional neural networks (CNNs) with shared weights and max pooling layers. We show that such CNNs produce linearly independent features at a “wide” layer which has more neurons than the number of training samples. This condition holds e.g. for the VGG network. Furthermore, we provide for such wide CNNs necessary and sufficient conditions for global minima with zero training error. For the case where the wide layer is followed by a fully connected layer we show that almost every critical point of the empirical loss is a global minimum with zero training error. Our analysis suggests that both depth and width are very important in deep learning. While depth brings more representational power and allows the network to learn high level features, width smoothes the optimization landscape of the loss function in the sense that a sufficiently wide network has a well-behaved loss surface with almost no bad local minima.
****
Neural Networks Should Be Wide Enough to Learn Disconnected Decision Regions
Quynh Nguyen, Mahesh Chandra Mukkamala, Matthias Hein
In the recent literature the important role of depth in deep learning has been emphasized. In this paper we argue that sufficient width of a feedforward network is equally important by answering the simple question under which conditions the decision regions of a neural network are connected. It turns out that for a class of activation functions including leaky ReLU, neural networks having a pyramidal structure, that is no layer has more hidden units than the input dimension, produce necessarily connected decision regions. This implies that a sufficiently wide hidden layer is necessary to guarantee that the network can produce disconnected decision regions. We discuss the implications of this result for the construction of neural networks, in particular the relation to the problem of adversarial manipulation of classifiers.
****
SGD and Hogwild! Convergence Without the Bounded Gradients Assumption
Lam Nguyen, PHUONG HA NGUYEN, Marten Dijk, Peter Richtarik, Katya Scheinberg, Martin Takac
Stochastic gradient descent (SGD) is the optimization algorithm of choice in many machine learning applications such as regularized empirical risk minimization and training deep neural networks. The classical convergence analysis of SGD is carried out under the assumption that the norm of the stochastic gradient is uniformly bounded. While this might hold for some loss functions, it is always violated for cases where the objective function is strongly convex. In (Bottou et al.,2016), a new analysis of convergence of SGD is performed under the assumption that stochastic gradients are bounded with respect to the true gradient norm. Here we show that for stochastic problems arising in machine learning such bound always holds; and we also propose an alternative convergence analysis of SGD with diminishing learning rate regime, which results in more relaxed conditions than those in (Bottou et al.,2016). We then move on the asynchronous parallel setting, and prove convergence of Hogwild! algorithm in the same regime, obtaining the first convergence results for this method in the case of diminished learning rate.
****
Active Testing: An Efficient and Robust Framework for Estimating Accuracy
Phuc Nguyen, Deva Ramanan, Charless Fowlkes
Much recent work on large-scale visual recogni- tion aims to scale up learning to massive, noisily- annotated datasets. We address the problem of scaling-up the evaluation of such models to large- scale datasets with noisy labels. Current protocols for doing so require a human user to either vet (re-annotate) a small fraction of the testset and ignore the rest, or else correct errors in annotation as they are found through manual inspection of results. In this work, we re-formulate the problem as one of active testing, and examine strategies for efficien

tly querying a user so as to obtain an accurate performance estimate with minima l vet- ting. We demonstrate the effectiveness of our proposed active testing fra mework on estimating two performance metrics, Precision@K and mean Average Preci sions, for two popular Computer Vi- sion tasks, multilabel classification and in stance segmentation, respectively. We further show that our approach is able to siginificantly save human annotation effort and more robust than alterna- tive e valuation protocols.
****

On Learning Sparsely Used Dictionaries from Incomplete Samples
Thanh Nguyen, Akshay Soni, Chinmay Hegde
Existing algorithms for dictionary learning assume that the entries of the (high -dimensional) input data are fully observed. However, in several practical appli cations, only an incomplete fraction of the data entries may be available. For i ncomplete settings, no provably correct and polynomial-time algorithm has been r eported in the dictionary learning literature. In this paper, we provide provabl e approaches for learning – from incomplete samples – a family of dictionaries w hose atoms have sufficiently “spread-out” mass. First, we propose a descent-styl e iterative algorithm that linearly converges to the true dictionary when provid ed a sufficiently coarse initial estimate. Second, we propose an initialization algorithm that utilizes a small number of extra fully observed samples to produc e such a coarse initial estimate. Finally, we theoretically analyze their perfor mance and provide asymptotic statistical and computational guarantees.
****

Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry
Maximillian Nickel, Douwe Kiela
We are concerned with the discovery of hierarchical relationships from large-sca le unstructured similarity scores. For this purpose, we study different models o f hyperbolic space and find that learning embeddings in the Lorentz model is sub stantially more efficient than in the Poincar{é}-ball model. We show that the pr oposed approach allows us to learn high-quality embeddings of large taxonomies w hich yield improvements over Poincar{é} embeddings, especially in low dimensions . Lastly, we apply our model to discover hierarchies in two real-world datasets: we show that an embedding in hyperbolic space can reveal important aspects of a company’s organizational structure as well as reveal historical relationships b etween language families.
****

State Space Gaussian Processes with Non-Gaussian Likelihood
Hannes Nickisch, Arno Solin, Alexander Grigorevskiy
We provide a comprehensive overview and tooling for GP modelling with non-Gaussi an likelihoods using state space methods. The state space formulation allows for solving one-dimensonal GP models in O(n) time and memory complexity. While exis ting literature has focused on the connection between GP regression and state sp ace methods, the computational primitives allowing for inference using general l ikelihoods in combination with the Laplace approximation (LA), variational Bayes (VB), and assumed density filtering (ADF) / expectation propagation (EP) scheme s has been largely overlooked. We present means of combining the efficient O(n) state space methodology with existing inference methods. We also furher extend e xisting methods, and provide unifying code implementing all approaches.
****

SparseMAP: Differentiable Sparse Structured Inference
Vlad Niculae, Andre Martins, Mathieu Blondel, Claire Cardie
Structured prediction requires searching over a combinatorial number of structur es. To tackle it, we introduce SparseMAP, a new method for sparse structured inf erence, together with corresponding loss functions. SparseMAP inference is able to automatically select only a few global structures: it is situated between MAP inference, which picks a single structure, and marginal inference, which assign s probability mass to all structures, including implausible ones. Importantly, S parseMAP can be computed using only calls to a MAP oracle, hence it is applicabl e even to problems where marginal inference is intractable, such as linear assig nment. Moreover, thanks to the solution sparsity, gradient backpropagation is ef

ficient regardless of the structure. SparseMAP thus enables us to augment deep n eural networks with generic and sparse structured hidden layers. Experiments in dependency parsing and natural language inference reveal competitive accuracy, i mproved interpretability, and the ability to capture natural language ambiguitie s, which is attractive for pipeline systems.
****
A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visu alizations
Weili Nie, Yang Zhang, Ankit Patel
Backpropagation-based visualizations have been proposed to interpret convolution al neural networks (CNNs), however a theory is missing to justify their behavior s: Guided backpropagation (GBP) and deconvolutional network (DeconvNet) generate more human-interpretable but less class-sensitive visualizations than saliency map. Motivated by this, we develop a theoretical explanation revealing that GBP and DeconvNet are essentially doing (partial) image recovery which is unrelated to the network decisions. Specifically, our analysis shows that the backward ReL U introduced by GBP and DeconvNet, and the local connections in CNNs are the two main causes of compelling visualizations. Extensive experiments are provided th at support the theoretical analysis.
****
Functional Gradient Boosting based on Residual Network Perception
Atsushi Nitanda, Taiji Suzuki
Residual Networks (ResNets) have become state-of-the-art models in deep learning and several theoretical studies have been devoted to understanding why ResNet w orks so well. One attractive viewpoint on ResNet is that it is optimizing the ri sk in a functional space by consisting of an ensemble of effective features. In this paper, we adopt this viewpoint to construct a new gradient boosting method, which is known to be very powerful in data analysis. To do so, we formalize the boosting perspective of ResNet mathematically using the notion of functional gr adients and propose a new method called ResFGB for classification tasks by lever aging ResNet perception. Two types of generalization guarantees are provided fro m the optimization perspective: one is the margin bound and the other is the exp ected risk bound by the sample-splitting technique. Experimental results show su perior performance of the proposed method over state-of-the-art methods such as LightGBM.
****
Beyond 1/2-Approximation for Submodular Maximization on Massive Data Streams
Ashkan Norouzi-Fard, Jakub Tarnawski, Slobodan Mitrovic, Amir Zandieh, Aidasadat Mousavifar, Ola Svensson
Many tasks in machine learning and data mining, such as data diversification, no n-parametric learning, kernel machines, clustering etc., require extracting a sm all but representative summary from a massive dataset. Often, such problems can be posed as maximizing a submodular set function subject to a cardinality constr aint. We consider this question in the streaming setting, where elements arrive over time at a fast pace and thus we need to design an efficient, low-memory alg orithm. One such method, proposed by Badanidiyuru et al. (2014), always finds a 0.5-approximate solution. Can this approximation factor be improved? We answer t his question affirmatively by designing a new algorithm Salsa for streaming subm odular maximization. It is the first low-memory, singlepass algorithm that impro ves the factor 0.5, under the natural assumption that elements arrive in a rando m order. We also show that this assumption is necessary, i.e., that there is no such algorithm with better than 0.5-approximation when elements arrive in arbitr ary order. Our experiments demonstrate that Salsa significantly outperforms the state of the art in applications related to exemplar-based clustering, social gr aph analysis, and recommender systems.
****
The Uncertainty Bellman Equation and Exploration
Brendan O’Donoghue, Ian Osband, Remi Munos, Vlad Mnih
We consider the exploration/exploitation problem in reinforcement learning. For exploitation, it is well known that the Bellman equation connects the value at a

ny time-step to the expected value at subsequent time-steps. In this paper we co
nsider a similar uncertainty Bellman equation (UBE), which connects the uncertai
nty at any time-step to the expected uncertainties at subsequent time-steps, the
reby extending the potential exploratory benefit of a policy beyond individual t
ime-steps. We prove that the unique fixed point of the UBE yields an upper bound
 on the variance of the posterior distribution of the Q-values induced by any po
licy. This bound can be much tighter than traditional count-based bonuses that c
ompound standard deviation rather than variance. Importantly, and unlike several
 existing approaches to optimism, this method scales naturally to large systems
with complex generalization. Substituting our UBE-exploration strategy for $\eps
ilon$-greedy improves DQN performance on 51 out of 57 games in the Atari suite.
****

Is Generator Conditioning Causally Related to GAN Performance?
Augustus Odena, Jacob Buckman, Catherine Olsson, Tom Brown, Christopher Olah, Co
lin Raffel, Ian Goodfellow
Recent work suggests that controlling the entire distribution of Jacobian singul
ar values is an important design consideration in deep learning. Motivated by th
is, we study the distribution of singular values of the Jacobian of the generato
r in Generative Adversarial Networks. We find that this Jacobian generally becom
es ill-conditioned at the beginning of training. Moreover, we find that the aver
age (across the latent space) conditioning of the generator is highly predictive
 of two other ad-hoc metrics for measuring the “quality” of trained GANs: the In
ception Score and the Frechet Inception Distance. We then test the hypothesis th
at this relationship is causal by proposing a “regularization” technique (called
 Jacobian Clamping) that softly penalizes the condition number of the generator
Jacobian. Jacobian Clamping improves the mean score for nearly all datasets on w
hich we tested it. It also greatly reduces inter-run variance of the aforementio
ned scores, addressing (at least partially) one of the main criticisms of GANs.
****

Learning in Reproducing Kernel Kre■■n Spaces
Dino Oglic, Thomas Gaertner
We formulate a novel regularized risk minimization problem for learning in repro
ducing kernel Kre{■■}n spaces and show that the strong representer theorem appli
es to it. As a result of the latter, the learning problem can be expressed as th
e minimization of a quadratic form over a hypersphere of constant radius. We pre
sent an algorithm that can find a globally optimal solution to this non-convex o
ptimization problem in time cubic in the number of instances. Moreover, we deriv
e the gradient of the solution with respect to its hyperparameters and, in this
way, provide means for efficient hyperparameter tuning. The approach comes with
a generalization bound expressed in terms of the Rademacher complexity of the co
rresponding hypothesis space. The major advantage over standard kernel methods i
s the ability to learn with various domain specific similarity measures for whic
h positive definiteness does not hold or is difficult to establish. The approach
 is evaluated empirically using indefinite kernels defined on structured as well
 as vectorial data. The empirical results demonstrate a superior performance of
our approach over the state-of-the-art baselines.
****

BOCK : Bayesian Optimization with Cylindrical Kernels
ChangYong Oh, Efstratios Gavves, Max Welling
A major challenge in Bayesian Optimization is the boundary issue where an algori
thm spends too many evaluations near the boundary of its search space. In this p
aper, we propose BOCK, Bayesian Optimization with Cylindrical Kernels, whose bas
ic idea is to transform the ball geometry of the search space using a cylindrica
l transformation. Because of the transformed geometry, the Gaussian Process-base
d surrogate model spends less budget searching near the boundary, while concentr
ating its efforts relatively more near the center of the search region, where we
 expect the solution to be located. We evaluate BOCK extensively, showing that i
t is not only more accurate and efficient, but it also scales successfully to pr
oblems with a dimensionality as high as 500. We show that the better accuracy an
d scalability of BOCK even allows optimizing modestly sized neural network layer

s, as well as neural network hyperparameters.
****
Self-Imitation Learning
Junhyuk Oh, Yijie Guo, Satinder Singh, Honglak Lee
This paper proposes Self-Imitation Learning (SIL), a simple off-policy actor-critic algorithm that learns to reproduce the agent's past good decisions. This algorithm is designed to verify our hypothesis that exploiting past good experiences can indirectly drive deep exploration. Our empirical results show that SIL significantly improves advantage actor-critic (A2C) on several hard exploration Atari games and is competitive to the state-of-the-art count-based exploration methods. We also show that SIL improves proximal policy optimization (PPO) on MuJoCo tasks.
****
A probabilistic framework for multi-view feature learning with many-to-many associations via neural networks
Akifumi Okuno, Tetsuya Hada, Hidetoshi Shimodaira
A simple framework Probabilistic Multi-view Graph Embedding (PMvGE) is proposed for multi-view feature learning with many-to-many associations so that it generalizes various existing multi-view methods. PMvGE is a probabilistic model for predicting new associations via graph embedding of the nodes of data vectors with links of their associations. Multi-view data vectors with many-to-many associations are transformed by neural networks to feature vectors in a shared space, and the probability of new association between two data vectors is modeled by the inner product of their feature vectors. While existing multi-view feature learning techniques can treat only either of many-to-many association or non-linear transformation, PMvGE can treat both simultaneously. By combining Mercer's theorem and the universal approximation theorem, we prove that PMvGE learns a wide class of similarity measures across views. Our likelihood-based estimator enables efficient computation of non-linear transformations of data vectors in large-scale datasets by minibatch SGD, and numerical experiments illustrate that PMvGE outperforms existing multi-view methods.
****
Transformation Autoregressive Networks
Junier Oliva, Avinava Dubey, Manzil Zaheer, Barnabas Poczos, Ruslan Salakhutdinov, Eric Xing, Jeff Schneider
The fundamental task of general density estimation $p(x)$ has been of keen interest to machine learning. In this work, we attempt to systematically characterize methods for density estimation. Broadly speaking, most of the existing methods can be categorized into either using: a) autoregressive models to estimate the conditional factors of the chain rule, $p(x_{i}\, |\, x_{i-1}, \ldots)$; or b) non-linear transformations of variables of a simple base distribution. Based on the study of the characteristics of these categories, we propose multiple novel methods for each category. For example we propose RNN based transformations to model non-Markovian dependencies. Further, through a comprehensive study over both real world and synthetic data, we show that jointly leveraging transformations of variables and autoregressive conditional models, results in a considerable improvement in performance. We illustrate the use of our models in outlier detection and image modeling. Finally we introduce a novel data driven framework for learning a family of distributions.
****
Design of Experiments for Model Discrimination Hybridising Analytical and Data-Driven Approaches
Simon Olofsson, Marc Deisenroth, Ruth Misener
Healthcare companies must submit pharmaceutical drugs or medical device to regulatory bodies before marketing new technology. Regulatory bodies frequently require transparent and interpretable computational modelling to justify a new healthcare technology, but researchers may have several competing models for a biological system and too little data to discriminate between the models. In design of experiments for model discrimination, where the goal is to design maximally informative physical experiments in order to discriminate between rival predictive m

odels, research has focused either on analytical approaches, which cannot manage all functions, or on data-driven approaches, which may have computational difficulties or lack interpretable marginal predictive distributions. We develop a methodology for introducing Gaussian process surrogates in lieu of the original mechanistic models. This allows us to extend existing design and model discrimination methods developed for analytical models to cases of non-analytical models.
****

Parallel WaveNet: Fast High-Fidelity Speech Synthesis
Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, Demis Hassabis
The recently-developed WaveNet architecture is the current state of the art in realistic speech synthesis, consistently rated as more natural sounding for many different languages than any previous system. However, because WaveNet relies on sequential generation of one audio sample at a time, it is poorly suited to today’s massively parallel computers, and therefore hard to deploy in a real-time production setting. This paper introduces Probability Density Distillation, a new method for training a parallel feed-forward network from a trained WaveNet with no significant difference in quality. The resulting system is capable of generating high-fidelity speech samples at more than 20 times faster than real-time, a 1000x speed up relative to the original WaveNet, and capable of serving multiple English and Japanese voices in a production setting.
****

Learning Localized Spatio-Temporal Models From Streaming Data
Muhammad Osama, Dave Zachariah, Thomas Schön
We address the problem of predicting spatio-temporal processes with temporal patterns that vary across spatial regions, when data is obtained as a stream. That is, when the training dataset is augmented sequentially. Specifically, we develop a localized spatio-temporal covariance model of the process that can capture spatially varying temporal periodicities in the data. We then apply a covariance-fitting methodology to learn the model parameters which yields a predictor that can be updated sequentially with each new data point. The proposed method is evaluated using both synthetic and real climate data which demonstrate its ability to accurately predict data missing in spatial regions over time.
****

Autoregressive Quantile Networks for Generative Modeling
Georg Ostrovski, Will Dabney, Remi Munos
We introduce autoregressive implicit quantile networks (AIQN), a fundamentally different approach to generative modeling than those commonly used, that implicitly captures the distribution using quantile regression. AIQN is able to achieve superior perceptual quality and improvements in evaluation metrics, without incurring a loss of sample diversity. The method can be applied to many existing models and architectures. In this work we extend the PixelCNN model with AIQN and demonstrate results on CIFAR-10 and ImageNet using Inception scores, FID, non-cherry-picked samples, and inpainting results. We consistently observe that AIQN yields a highly stable algorithm that improves perceptual quality while maintaining a highly diverse distribution.
****

Efficient First-Order Algorithms for Adaptive Signal Denoising
Dmitrii Ostrovskii, Zaid Harchaoui
We consider the problem of discrete-time signal denoising, focusing on a specific family of non-linear convolution-type estimators. Each such estimator is associated with a time-invariant filter which is obtained adaptively, by solving a certain convex optimization problem. Adaptive convolution-type estimators were demonstrated to have favorable statistical properties, see (Juditsky & Nemirovski, 2009; 2010; Harchaoui et al., 2015b; Ostrovsky et al., 2016). Our first contribution is an efficient implementation of these estimators via the known first-order proximal algorithms. Our second contribution is a computational complexity analysis of the proposed procedures, which takes into account their statistical nat

ure and the related notion of statistical accuracy. The proposed procedures and their analysis are illustrated on a simulated data benchmark.
****

Analyzing Uncertainty in Neural Machine Translation
Myle Ott, Michael Auli, David Grangier, Marc'Aurelio Ranzato
Machine translation is a popular test bed for research in neural sequence-to-sequence models but despite much recent research, there is still a lack of understanding of these models. Practitioners report performance degradation with large beams, the under-estimation of rare words and a lack of diversity in the final translations. Our study relates some of these issues to the inherent uncertainty of the task, due to the existence of multiple valid translations for a single source sentence, and to the extrinsic uncertainty caused by noisy training data. We propose tools and metrics to assess how uncertainty in the data is captured by the model distribution and how it affects search strategies that generate translations. Our results show that search works remarkably well but that the models tend to spread too much probability mass over the hypothesis space. Next, we propose tools to assess model calibration and show how to easily fix some shortcomings of current models. We release both code and multiple human reference translations for two popular benchmarks.
****

Learning Compact Neural Networks with Regularization
Samet Oymak
Proper regularization is critical for speeding up training, improving generalization performance, and learning compact models that are cost efficient. We propose and analyze regularized gradient descent algorithms for learning shallow neural networks. Our framework is general and covers weight-sharing (convolutional networks), sparsity (network pruning), and low-rank constraints among others. We first introduce covering dimension to quantify the complexity of the constraint set and provide insights on the generalization properties. Then, we show that proposed algorithms become well-behaved and local linear convergence occurs once the amount of data exceeds the covering dimension. Overall, our results demonstrate that near-optimal sample complexity is sufficient for efficient learning and illustrate how regularization can be beneficial to learn over-parameterized networks.
****

Tree Edit Distance Learning via Adaptive Symbol Embeddings
Benjamin Paaßen, Claudio Gallicchio, Alessio Micheli, Barbara Hammer
Metric learning has the aim to improve classification accuracy by learning a distance measure which brings data points from the same class closer together and pushes data points from different classes further apart. Recent research has demonstrated that metric learning approaches can also be applied to trees, such as molecular structures, abstract syntax trees of computer programs, or syntax trees of natural language, by learning the cost function of an edit distance, i.e. the costs of replacing, deleting, or inserting nodes in a tree. However, learning such costs directly may yield an edit distance which violates metric axioms, is challenging to interpret, and may not generalize well. In this contribution, we propose a novel metric learning approach for trees which we call embedding edit distance learning (BEDL) and which learns an edit distance indirectly by embedding the tree nodes as vectors, such that the Euclidean distance between those vectors supports class discrimination. We learn such embeddings by reducing the distance to prototypical trees from the same class and increasing the distance to prototypical trees from different classes. In our experiments, we show that BEDL improves upon the state-of-the-art in metric learning for trees on six benchmark data sets, ranging from computer science over biomedical data to a natural-language processing data set containing over 300,000 nodes.
****

Reinforcement Learning with Function-Valued Action Spaces for Partial Differential Equation Control
Yangchen Pan, Amir-massoud Farahmand, Martha White, Saleh Nabi, Piyush Grover, Daniel Nikovski

Recent work has shown that reinforcement learning (RL) is a promising approach t
o control dynamical systems described by partial differential equations (PDE). T
his paper shows how to use RL to tackle more general PDE control problems that h
ave continuous high-dimensional action spaces with spatial relationship among ac
tion dimensions. In particular, we propose the concept of action descriptors, wh
ich encode regularities among spatially-extended action dimensions and enable th
e agent to control high-dimensional action PDEs. We provide theoretical evidence
 suggesting that this approach can be more sample efficient compared to a conven
tional approach that treats each action dimension separately and does not explic
itly exploit the spatial regularity of the action space. The action descriptor a
pproach is then used within the deep deterministic policy gradient algorithm. Ex
periments on two PDE control problems, with up to 256-dimensional continuous act
ions, show the advantage of the proposed approach over the conventional one.
****

Learning to Speed Up Structured Output Prediction
Xingyuan Pan, Vivek Srikumar
Predicting structured outputs can be computationally onerous due to the combinat
orially large output spaces. In this paper, we focus on reducing the prediction
time of a trained black-box structured classifier without losing accuracy. To do
 so, we train a speedup classifier that learns to mimic a black-box classifier u
nder the learning-to-search approach. As the structured classifier predicts more
 examples, the speedup classifier will operate as a learned heuristic to guide s
earch to favorable regions of the output space. We present a mistake bound for t
he speedup classifier and identify inference situations where it can independent
ly make correct judgments without input features. We evaluate our method on the
task of entity and relation extraction and show that the speedup classifier outp
erforms even greedy search in terms of speed without loss of accuracy.
****

Theoretical Analysis of Image-to-Image Translation with Adversarial Learning
Xudong Pan, Mi Zhang, Daizong Ding
Recently, a unified model for image-to-image translation tasks within adversaria
l learning framework has aroused widespread research interests in computer visio
n practitioners. Their reported empirical success however lacks solid theoretica
l interpretations for its inherent mechanism. In this paper, we reformulate thei
r model from a brand-new geometrical perspective and have eventually reached a f
ull interpretation on some interesting but unclear empirical phenomenons from th
eir experiments. Furthermore, by extending the definition of generalization for
generative adversarial nets to a broader sense, we have derived a condition to c
ontrol the generalization capability of their model. According to our derived co
ndition, several practical suggestions have also been proposed on model design a
nd dataset construction as a guidance for further empirical researches.
****

Max-Mahalanobis Linear Discriminant Analysis Networks
Tianyu Pang, Chao Du, Jun Zhu
A deep neural network (DNN) consists of a nonlinear transformation from an input
 to a feature representation, followed by a common softmax linear classifier. Th
ough many efforts have been devoted to designing a proper architecture for nonli
near transformation, little investigation has been done on the classifier part.
In this paper, we show that a properly designed classifier can improve robustnes
s to adversarial attacks and lead to better prediction results. Specifically, we
 define a Max-Mahalanobis distribution (MMD) and theoretically show that if the
input distributes as a MMD, the linear discriminant analysis (LDA) classifier wi
ll have the best robustness to adversarial examples. We further propose a novel
Max-Mahalanobis linear discriminant analysis (MM-LDA) network, which explicitly
maps a complicated data distribution in the input space to a MMD in the latent f
eature space and then applies LDA to make predictions. Our results demonstrate t
hat the MM-LDA networks are significantly more robust to adversarial attacks, an
d have better performance in class-biased classification.
****

Stochastic Variance-Reduced Policy Gradient

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, Marcello Restelli

In this paper, we propose a novel reinforcement-learning algorithm consisting in a stochastic variance-reduced version of policy gradient for solving Markov Decision Processes (MDPs). Stochastic variance-reduced gradient (SVRG) methods have proven to be very successful in supervised learning. However, their adaptation to policy gradient is not straightforward and needs to account for I) a non-concave objective function; II) approximations in the full gradient computation; and III) a non-stationary sampling process. The result is SVRPG, a stochastic variance-reduced policy gradient algorithm that leverages on importance weights to preserve the unbiasedness of the gradient estimate. Under standard assumptions on the MDP, we provide convergence guarantees for SVRPG with a convergence rate that is linear under increasing batch sizes. Finally, we suggest practical variants of SVRPG, and we empirically evaluate them on continuous MDPs.

****

Learning Independent Causal Mechanisms

Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, Bernhard Schölkopf

Statistical learning relies upon data sampled from a distribution, and we usually do not care what actually generated it in the first place. From the point of view of causal modeling, the structure of each distribution is induced by physical mechanisms that give rise to dependences between observables. Mechanisms, however, can be meaningful autonomous modules of generative models that make sense beyond a particular entailed data distribution, lending themselves to transfer between problems. We develop an algorithm to recover a set of independent (inverse) mechanisms from a set of transformed data points. The approach is unsupervised and based on a set of experts that compete for data generated by the mechanisms, driving specialization. We analyze the proposed method in a series of experiments on image data. Each expert learns to map a subset of the transformed data back to a reference distribution. The learned mechanisms generalize to novel domains. We discuss implications for transfer learning and links to recent trends in generative modeling.

****

Time Limits in Reinforcement Learning

Fabio Pardo, Arash Tavakoli, Vitaly Levdik, Petar Kormushev

In reinforcement learning, it is common to let an agent interact for a fixed amount of time with its environment before resetting it and repeating the process in a series of episodes. The task that the agent has to learn can either be to maximize its performance over (i) that fixed period, or (ii) an indefinite period where time limits are only used during training to diversify experience. In this paper, we provide a formal account for how time limits could effectively be handled in each of the two cases and explain why not doing so can cause state-aliasing and invalidation of experience replay, leading to suboptimal policies and training instability. In case (i), we argue that the terminations due to time limits are in fact part of the environment, and thus a notion of the remaining time should be included as part of the agent’s input to avoid violation of the Markov property. In case (ii), the time limits are not part of the environment and are only used to facilitate learning. We argue that this insight should be incorporated by bootstrapping from the value of the state at the end of each partial episode. For both cases, we illustrate empirically the significance of our considerations in improving the performance and stability of existing reinforcement learning algorithms, showing state-of-the-art results on several control tasks.

****

Image Transformer

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, Dustin Tran

Image generation has been successfully cast as an autoregressive sequence generation or transformation problem. Recent work has shown that self-attention is an effective way of modeling textual sequences. In this work, we generalize a recently proposed model architecture based on self-attention, the Transformer, to a s

equence modeling formulation of image generation with a tractable likelihood. By restricting the self-attention mechanism to attend to local neighborhoods we significantly increase the size of images the model can process in practice, despite maintaining significantly larger receptive fields per layer than typical convolutional neural networks. While conceptually simple, our generative models significantly outperform the current state of the art in image generation on ImageNet, improving the best published negative log-likelihood on ImageNet from 3.83 to 3.77. We also present results on image super-resolution with a large magnification ratio, applying an encoder-decoder configuration of our architecture. In a human evaluation study, we find that images generated by our super-resolution model fool human observers three times more often than the previous state of the art.

****

PIPPS: Flexible Model-Based Policy Search Robust to the Curse of Chaos

Paavo Parmas, Carl Edward Rasmussen, Jan Peters, Kenji Doya

Previously, the exploding gradient problem has been explained to be central in deep learning and model-based reinforcement learning, because it causes numerical issues and instability in optimization. Our experiments in model-based reinforcement learning imply that the problem is not just a numerical issue, but it may be caused by a fundamental chaos-like nature of long chains of nonlinear computations. Not only do the magnitudes of the gradients become large, the direction of the gradients becomes essentially random. We show that reparameterization gradients suffer from the problem, while likelihood ratio gradients are robust. Using our insights, we develop a model-based policy search framework, Probabilistic Inference for Particle-Based Policy Search (PIPPS), which is easily extensible, and allows for almost arbitrary models and policies, while simultaneously matching the performance of previous data-efficient learning algorithms. Finally, we invent the total propagation algorithm, which efficiently computes a union over all pathwise derivative depths during a single backwards pass, automatically giving greater weight to estimators with lower variance, sometimes improving over reparameterization gradients by 10^6 times.

****

High-Quality Prediction Intervals for Deep Learning: A Distribution-Free, Ensembled Approach

Tim Pearce, Alexandra Brintrup, Mohamed Zaki, Andy Neely

This paper considers the generation of prediction intervals (PIs) by neural networks for quantifying uncertainty in regression tasks. It is axiomatic that high-quality PIs should be as narrow as possible, whilst capturing a specified portion of data. We derive a loss function directly from this axiom that requires no distributional assumption. We show how its form derives from a likelihood principle, that it can be used with gradient descent, and that model uncertainty is accounted for in ensembled form. Benchmark experiments show the method outperforms current state-of-the-art uncertainty quantification methods, reducing average PI width by over 10%.

****

Adaptive Three Operator Splitting

Fabian Pedregosa, Gauthier Gidel

We propose and analyze a novel adaptive step size variant of the Davis-Yin three operator splitting, a method that can solve optimization problems composed of a sum of a smooth term for which we have access to its gradient and an arbitrary number of potentially non-smooth terms for which we have access to their proximal operator. The proposed method leverages local information of the objective function, allowing for larger step sizes while preserving the convergence properties of the original method. It only requires two extra function evaluations per iteration and does not depend on any step size hyperparameter besides an initial estimate. We provide a convergence rate analysis of this method, showing sublinear convergence rate for general convex functions and linear convergence under stronger assumptions, matching the best known rates of its non adaptive variant. Finally, an empirical comparison with related methods on 6 different problems illustrates the computational advantage of the adaptive step size strategy.

****
Efficient Neural Architecture Search via Parameters Sharing
Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, Jeff Dean
We propose Efficient Neural Architecture Search (ENAS), a fast and inexpensive approach for automatic model design. ENAS constructs a large computational graph, where each subgraph represents a neural network architecture, hence forcing all architectures to share their parameters. A controller is trained with policy gradient to search for a subgraph that maximizes the expected reward on a validation set. Meanwhile a model corresponding to the selected subgraph is trained to minimize a canonical cross entropy loss. Sharing parameters among child models allows ENAS to deliver strong empirical performances, whilst using much fewer GPU-hours than existing automatic model design approaches, and notably, 1000x less expensive than standard Neural Architecture Search. On Penn Treebank, ENAS discovers a novel architecture that achieves a test perplexity of 56.3, on par with the existing state-of-the-art among all methods without post-training processing. On CIFAR-10, ENAS finds a novel architecture that achieves 2.89% test error, which is on par with the 2.65% test error of NASNet (Zoph et al., 2018).
****
Bandits with Delayed, Aggregated Anonymous Feedback
Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, Steffen Grunewalder
We study a variant of the stochastic K-armed bandit problem, which we call "bandits with delayed, aggregated anonymous feedback”. In this problem, when the player pulls an arm, a reward is generated, however it is not immediately observed. Instead, at the end of each round the player observes only the sum of a number of previously generated rewards which happen to arrive in the given round. The rewards are stochastically delayed and due to the aggregated nature of the observations, the information of which arm led to a particular reward is lost. The question is what is the cost of the information loss due to this delayed, aggregated anonymous feedback? Previous works have studied bandits with stochastic, non-anonymous delays and found that the regret increases only by an additive factor relating to the expected delay. In this paper, we show that this additive regret increase can be maintained in the harder delayed, aggregated anonymous feedback setting when the expected delay (or a bound on it) is known. We provide an algorithm that matches the worst case regret of the non-anonymous problem exactly when the delays are bounded, and up to logarithmic factors or an additive variance term for unbounded delays.
****
Constant-Time Predictive Distributions for Gaussian Processes
Geoff Pleiss, Jacob Gardner, Kilian Weinberger, Andrew Gordon Wilson
One of the most compelling features of Gaussian process (GP) regression is its ability to provide well-calibrated posterior distributions. Recent advances in inducing point methods have sped up GP marginal likelihood and posterior mean computations, leaving posterior covariance estimation and sampling as the remaining computational bottlenecks. In this paper we address these shortcomings by using the Lanczos algorithm to rapidly approximate the predictive covariance matrix. Our approach, which we refer to as LOVE (LanczOs Variance Estimates), substantially improves time and space complexity. In our experiments, LOVE computes covariances up to 2,000 times faster and draws samples 18,000 times faster than existing methods, all without sacrificing accuracy.
****
Local Convergence Properties of SAGA/Prox-SVRG and Acceleration
Clarice Poon, Jingwei Liang, Carola Schoenlieb
In this paper, we present a local convergence anal- ysis for a class of stochastic optimisation meth- ods: the proximal variance reduced stochastic gradient methods, and mainly focus on SAGA (Defazio et al., 2014) and Prox-SVRG (Xiao & Zhang, 2014). Under the assumption that the non-smooth component of the optimisation prob- lem is partly smooth relative to a smooth mani- fold, we present a unified framework for the local convergence analysis of SAGA/Prox-SVRG: (i) the sequences generated by the methods are able to identify the smooth manifold in a finite num- ber of iterations; (ii) then the sequence enters a local linear convergen

ce regime. Furthermore, we discuss various possibilities for accelerating these algorithms, including adapting to better lo- cal parameters, and applying higher -order deter- ministic/stochastic optimisation methods which can achieve super-l inear convergence. Several concrete examples arising from machine learning are c onsidered to demonstrate the obtained result.
****

Equivalence of Multicategory SVM and Simplex Cone SVM: Fast Computations and Sta tistical Theory
Guillaume Pouliot
The multicategory SVM (MSVM) of Lee et al. (2004) is a natural generalization of the classical, binary support vector machines (SVM). However, its use has been limited by computational difficulties. The simplex-cone SVM (SCSVM) of Mroueh et al. (2012) is a computationally efficient multicategory classifier, but its use has been limited by a seemingly opaque interpretation. We show that MSVM and SC SVM are in fact exactly equivalent, and provide a bijection between their tuning parameters. MSVM may then be entertained as both a natural and computationally efficient multicategory extension of SVM. We further provide a Donsker theorem f or finite-dimensional kernel MSVM and partially answer the open question pertain ing to the very competitive performance of One-vs-Rest methods against MSVM. Fur thermore, we use the derived asymptotic covariance formula to develop an inverse -variance weighted classification rule which improves on the One-vs-Rest approac h.
****

Learning Dynamics of Linear Denoising Autoencoders
Arnu Pretorius, Steve Kroon, Herman Kamper
Denoising autoencoders (DAEs) have proven useful for unsupervised representation learning, but a thorough theoretical understanding is still lacking of how the input noise influences learning. Here we develop theory for how noise influences learning in DAEs. By focusing on linear DAEs, we are able to derive analytic ex pressions that exactly describe their learning dynamics. We verify our theoretic al predictions with simulations as well as experiments on MNIST and CIFAR-10. Th e theory illustrates how, when tuned correctly, noise allows DAEs to ignore low variance directions in the inputs while learning to reconstruct them. Furthermor e, in a comparison of the learning dynamics of DAEs to standard regularised auto encoders, we show that noise has a similar regularisation effect to weight decay , but with faster training dynamics. We also show that our theoretical predictio ns approximate learning dynamics on real-world data and qualitatively match obse rved dynamics in nonlinear DAEs.
****

JointGAN: Multi-Domain Joint Distribution Learning with Generative Adversarial N ets
Yunchen Pu, Shuyang Dai, Zhe Gan, Weiyao Wang, Guoyin Wang, Yizhe Zhang, Ricardo Henao, Lawrence Carin Duke
A new generative adversarial network is developed for joint distribution matchin g.Distinct from most existing approaches, that only learn conditional distributi ons, the proposed model aims to learn a joint distribution of multiple random va riables (domains). This is achieved by learning to sample from conditional distr ibutions between the domains, while simultaneously learning to sample from the m arginals of each individual domain.The proposed framework consists of multiple g enerators and a single softmax-based critic, all jointly trained via adversarial learning.From a simple noise source, the proposed framework allows synthesis of draws from the marginals, conditional draws given observations from a subset of random variables, or complete draws from the full joint distribution. Most exam ples considered are for joint analysis of two domains, with examples for three d omains also presented.
****

Selecting Representative Examples for Program Synthesis
Yewen Pu, Zachery Miranda, Armando Solar-Lezama, Leslie Kaelbling
Program synthesis is a class of regression problems where one seeks a solution, in the form of a source-code program, mapping the inputs to their corresponding

outputs exactly. Due to its precise and combinatorial nature, program synthesis is commonly formulated as a constraint satisfaction problem, where input-output examples are encoded as constraints and solved with a constraint solver. A key c hallenge of this formulation is scalability: while constraint solvers work well with a few well-chosen examples, a large set of examples can incur significant o verhead in both time and memory. We describe a method to discover a subset of ex amples that is both small and representative: the subset is constructed iterativ ely, using a neural network to predict the probability of unchosen examples cond itioned on the chosen examples in the subset, and greedily adding the least prob able example. We empirically evaluate the representativeness of the subsets cons tructed by our method, and demonstrate such subsets can significantly improve sy nthesis time and stability.

****

Generalized Earley Parser: Bridging Symbolic Grammars and Sequence Data for Futu re Prediction

Siyuan Qi, Baoxiong Jia, Song-Chun Zhu

Future predictions on sequence data (e.g., videos or audios) require the algorit hms to capture non-Markovian and compositional properties of high-level semantic s. Context-free grammars are natural choices to capture such properties, but tra ditional grammar parsers (e.g., Earley parser) only take symbolic sentences as i nputs. In this paper, we generalize the Earley parser to parse sequence data whi ch is neither segmented nor labeled. This generalized Earley parser integrates a grammar parser with a classifier to find the optimal segmentation and labels, a nd makes top-down future predictions. Experiments show that our method significa ntly outperforms other approaches for future human activity prediction.

****

Do Outliers Ruin Collaboration?

Mingda Qiao

We consider the problem of learning a binary classifier from n different data sources, among which at most an η fraction are adversarial. The overhead is defined as the ratio between the sample complexity of learning in this setting and that of learning the same hypothesis class on a single data distribution. We present an algorithm that achieves an $O(\eta n + \ln n)$ overhead, which is pr oved to be worst-case optimal. We also discuss the potential challenges to the d esign of a computationally efficient learning algorithm with a small overhead.

****

Gradually Updated Neural Networks for Large-Scale Image Recognition

Siyuan Qiao, Zhishuai Zhang, Wei Shen, Bo Wang, Alan Yuille

Depth is one of the keys that make neural networks succeed in the task of large- scale image recognition. The state-of-the-art network architectures usually incr ease the depths by cascading convolutional layers or building blocks. In this pa per, we present an alternative method to increase the depth. Our method is by in troducing computation orderings to the channels within convolutional layers or b locks, based on which we gradually compute the outputs in a channel-wise manner. The added orderings not only increase the depths and the learning capacities of the networks without any additional computation costs, but also eliminate the o verlap singularities so that the networks are able to converge faster and perfor m better. Experiments show that the networks based on our method achieve the sta te-of-the-art performances on CIFAR and ImageNet datasets.

****

DCFNet: Deep Neural Network with Decomposed Convolutional Filters

Qiang Qiu, Xiuyuan Cheng, Calderbank, Guillermo Sapiro

Filters in a Convolutional Neural Network (CNN) contain model parameters learned from enormous amounts of data. In this paper, we suggest to decompose convoluti onal filters in CNN as a truncated expansion with pre-fixed bases, namely the De composed Convolutional Filters network (DCFNet), where the expansion coefficient s remain learned from data. Such a structure not only reduces the number of trai nable parameters and computation, but also imposes filter regularity by bases tr uncation. Through extensive experiments, we consistently observe that DCFNet mai ntains accuracy for image classification tasks with a significant reduction of m

odel parameters, particularly with Fourier-Bessel (FB) bases, and even with rand
om bases. Theoretically, we analyze the representation stability of DCFNet with
respect to input variations, and prove representation stability under generic as
sumptions on the expansion coefficients. The analysis is consistent with the emp
irical observations.
****

Non-convex Conditional Gradient Sliding
Chao Qu, Yan Li, Huan Xu
We investigate a projection free optimization method, namely non-convex conditio
nal gradient sliding (NCGS) for non-convex optimization problems on the batch, s
tochastic and finite-sum settings. Conditional gradient sliding (CGS) method, by
 integrating Nesterov’s accelerated gradient method with Frank-Wolfe (FW) method
 in a smart way, outperforms FW for convex optimization, by reducing the amount
of gradient computations. However, the study of CGS in the non-convex setting is
 limited. In this paper, we propose the non-convex conditional gradient sliding
(NCGS) methods and analyze their convergence properties. We also leverage the id
ea of variance reduction from the recent progress in convex optimization to obta
in a new algorithm termed variance reduced NCGS (NCGS-VR), and obtain faster con
vergence rate than the batch NCGS in the finite-sum setting. We show that NCGS a
lgorithms outperform their Frank-Wolfe counterparts both in theory and in practi
ce, for all three settings, namely the batch, stochastic and finite-sum setting.
 This significantly improves our understanding of optimizing non-convex function
s with complicated feasible sets (where projection is prohibitively expensive).
****

Machine Theory of Mind
Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, Ma
tthew Botvinick
Theory of mind (ToM) broadly refers to humans’ ability to represent the mental s
tates of others, including their desires, beliefs, and intentions. We design a T
heory of Mind neural network {–} a ToMnet {–} which uses meta-learning to build
such models of the agents it encounters. The ToMnet learns a strong prior model
for agents’ future behaviour, and, using only a small number of behavioural obse
rvations, can bootstrap to richer predictions about agents’ characteristics and
mental states. We apply the ToMnet to agents behaving in simple gridworld enviro
nments, showing that it learns to model random, algorithmic, and deep RL agents
from varied populations, and that it passes classic ToM tasks such as the "Sally
-Anne" test of recognising that others can hold false beliefs about the world.
****

Fast Parametric Learning with Activation Memorization
Jack Rae, Chris Dyer, Peter Dayan, Timothy Lillicrap
Neural networks trained with backpropagation often struggle to identify classes
that have been observed a small number of times. In applications where most clas
s labels are rare, such as language modelling, this can become a performance bot
tleneck. One potential remedy is to augment the network with a fast-learning non
-parametric model which stores recent activations and class labels into an exter
nal memory. We explore a simplified architecture where we treat a subset of the
model parameters as fast memory stores. This can help retain information over lo
nger time intervals than a traditional memory, and does not require additional s
pace or compute. In the case of image classification, we display faster binding
of novel classes on an Omniglot image curriculum task. We also show improved per
formance for word-based language models on news reports (GigaWord), books (Proje
ct Gutenberg) and Wikipedia articles (WikiText-103) - the latter achieving a sta
te-of-the-art perplexity of 29.2.
****

Can Deep Reinforcement Learning Solve Erdos-Selfridge-Spencer Games?
Maithra Raghu, Alex Irpan, Jacob Andreas, Bobby Kleinberg, Quoc Le, Jon Kleinber
g
Deep reinforcement learning has achieved many recent successes, but our understa
nding of its strengths and limitations is hampered by the lack of rich environme
nts in which we can fully characterize optimal behavior, and correspondingly dia

gnose individual actions against such a characterization. Here we consider a family of combinatorial games, arising from work of Erdos, Selfridge, and Spencer, and we propose their use as environments for evaluating and comparing different approaches to reinforcement learning. These games have a number of appealing features: they are challenging for current learning approaches, but they form (i) a low-dimensional, simply parametrized environment where (ii) there is a linear closed form solution for optimal behavior from any state, and (iii) the difficulty of the game can be tuned by changing environment parameters in an interpretable way. We use these Erdos-Selfridge-Spencer games not only to compare different algorithms, but test for generalization, make comparisons to supervised learning, analyse multiagent play, and even develop a self play algorithm.

****

Cut-Pursuit Algorithm for Regularizing Nonsmooth Functionals with Graph Total Variation

Hugo Raguet, Loic Landrieu

We present an extension of the cut-pursuit algorithm, introduced by Landrieu and Obozinski (2017), to the graph total-variation regularization of functions with a separable nondifferentiable part. We propose a modified algorithmic scheme as well as adapted proofs of convergence. We also present a heuristic approach for handling the cases in which the values associated to each vertex of the graph are multidimensional. The performance of our algorithm, which we demonstrate on difficult, ill-conditioned large-scale inverse and learning problems, is such that it may in practice extend the scope of application of the total-variation regularization.

****

Modeling Others using Oneself in Multi-Agent Reinforcement Learning

Roberta Raileanu, Emily Denton, Arthur Szlam, Rob Fergus

We consider the multi-agent reinforcement learning setting with imperfect information. The reward function depends on the hidden goals of both agents, so the agents must infer the other players' goals from their observed behavior in order to maximize their returns. We propose a new approach for learning in these domains: Self Other-Modeling (SOM), in which an agent uses its own policy to predict the other agent's actions and update its belief of their hidden goal in an online manner. We evaluate this approach on three different tasks and show that the agents are able to learn better policies using their estimate of the other players' goals, in both cooperative and competitive settings.

****

On Nesting Monte Carlo Estimators

Tom Rainforth, Rob Cornish, Hongseok Yang, Andrew Warrington, Frank Wood

Many problems in machine learning and statistics involve nested expectations and thus do not permit conventional Monte Carlo (MC) estimation. For such problems, one must nest estimators, such that terms in an outer estimator themselves involve calculation of a separate, nested, estimation. We investigate the statistical implications of nesting MC estimators, including cases of multiple levels of nesting, and establish the conditions under which they converge. We derive corresponding rates of convergence and provide empirical evidence that these rates are observed in practice. We further establish a number of pitfalls that can arise from naive nesting of MC estimators, provide guidelines about how these can be avoided, and lay out novel methods for reformulating certain classes of nested expectation problems into single expectations, leading to improved convergence rates. We demonstrate the applicability of our work by using our results to develop a new estimator for discrete Bayesian experimental design problems and derive error bounds for a class of variational objectives.

****

Tighter Variational Bounds are Not Necessarily Better

Tom Rainforth, Adam Kosiorek, Tuan Anh Le, Chris Maddison, Maximilian Igl, Frank Wood, Yee Whye Teh

We provide theoretical and empirical evidence that using tighter evidence lower bounds (ELBOs) can be detrimental to the process of learning an inference network by reducing the signal-to-noise ratio of the gradient estimator. Our results c

all into question common implicit assumptions that tighter ELBOs are better vari
ational objectives for simultaneous model learning and inference amortization sc
hemes. Based on our insights, we introduce three new algorithms: the partially i
mportance weighted auto-encoder (PIWAE), the multiply importance weighted auto-e
ncoder (MIWAE), and the combination importance weighted autoencoder (CIWAE), eac
h of which includes the standard importance weighted auto-encoder (IWAE) as a sp
ecial case. We show that each can deliver improvements over IWAE, even when perf
ormance is measured by the IWAE target itself. Furthermore, our results suggest
that PIWAE may be able to deliver simultaneous improvements in the training of b
oth the inference and generative networks.
****

SAFFRON: an Adaptive Algorithm for Online Control of the False Discovery Rate
Aaditya Ramdas, Tijana Zrnic, Martin Wainwright, Michael Jordan
In the online false discovery rate (FDR) problem, one observes a possibly infini
te sequence of p-values $P_1,P_2,…$, each testing a different null hypothesis,
 and an algorithm must pick a sequence of rejection thresholds $\alpha_1,\alpha_
2,…$ in an online fashion, effectively rejecting the k-th null hypothesis when
ever $P_k \leq \alpha_k$. Importantly, α_k must be a function of the past
, and cannot depend on P_k or any of the later unseen p-values, and must be
chosen to guarantee that for any time t, the FDR up to time t is less than s
ome pre-determined quantity $\alpha \in (0,1)$. In this work, we present a power
ful new framework for online FDR control that we refer to as “SAFFRON”. Like old
er alpha-investing algorithms, SAFFRON starts off with an error budget (called a
lpha-wealth) that it intelligently allocates to different tests over time, earni
ng back some alpha-wealth whenever it makes a new discovery. However, unlike old
er methods, SAFFRON’s threshold sequence is based on a novel estimate of the alp
ha fraction that it allocates to true null hypotheses. In the offline setting, a
lgorithms that employ an estimate of the proportion of true nulls are called “ad
aptive”, hence SAFFRON can be seen as an online analogue of the offline Storey-B
H adaptive procedure. Just as Storey-BH is typically more powerful than the Benj
amini-Hochberg (BH) procedure under independence, we demonstrate that SAFFRON is
 also more powerful than its non-adaptive counterparts such as LORD.
****

QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement
Learning
Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob F
oerster, Shimon Whiteson
In many real-world settings, a team of agents must coordinate their behaviour wh
ile acting in a decentralised way. At the same time, it is often possible to tra
in the agents in a centralised fashion in a simulated or laboratory setting, whe
re global state information is available and communication constraints are lifte
d. Learning joint action-values conditioned on extra state information is an att
ractive way to exploit centralised learning, but the best strategy for then extr
acting decentralised policies is unclear. Our solution is QMIX, a novel value-ba
sed method that can train decentralised policies in a centralised end-to-end fas
hion. QMIX employs a network that estimates joint action-values as a complex non
-linear combination of per-agent values that condition only on local observation
s. We structurally enforce that the joint-action value is monotonic in the per-a
gent values, which allows tractable maximisation of the joint action-value in of
f-policy learning, and guarantees consistency between the centralised and decent
ralised policies. We evaluate QMIX on a challenging set of StarCraft II microman
agement tasks, and show that QMIX significantly outperforms existing value-based
 multi-agent reinforcement learning methods.
****

Gradient Coding from Cyclic MDS Codes and Expander Graphs
Netanel Raviv, Rashish Tandon, Alex Dimakis, Itzhak Tamo
Gradient coding is a technique for straggler mitigation in distributed learning.
 In this paper we design novel gradient codes using tools from classical coding
theory, namely, cyclic MDS codes, which compare favourably with existing solutio
ns, both in the applicable range of parameters and in the complexity of the invo

lved algorithms. Second, we introduce an approximate variant of the gradient cod ing problem, in which we settle for approximate gradient computation instead of the exact one. This approach enables graceful degradation, i.e., the ℓ_2 er ror of the approximate gradient is a decreasing function of the number of stragg lers. Our main result is that the normalized adjacency matrix of an expander gra ph can yield excellent approximate gradient codes, and that this approach allows us to perform significantly less computation compared to exact gradient coding. We experimentally test our approach on Amazon EC2, and show that the generaliza tion error of approximate gradient coding is very close to the full gradient whi le requiring significantly less computation from the workers.
****

Learning Implicit Generative Models with the Method of Learned Moments
Suman Ravuri, Shakir Mohamed, Mihaela Rosca, Oriol Vinyals
We propose a method of moments (MoM) algorithm for training large-scale implicit generative models. Moment estimation in this setting encounters two problems: i t is often difficult to define the millions of moments needed to learn the model parameters, and it is hard to determine which properties are useful when specif ying moments. To address the first issue, we introduce a moment network, and def ine the moments as the network’s hidden units and the gradient of the network’s output with respect to its parameters. To tackle the second problem, we use asym ptotic theory to highlight desiderata for moments – namely they should minimize the asymptotic variance of estimated model parameters – and introduce an objecti ve to learn better moments. The sequence of objectives created by this Method of Learned Moments (MoLM) can train high-quality neural image samplers. On CIFAR-1 0, we demonstrate that MoLM-trained generators achieve significantly higher Ince ption Scores and lower Frechet Inception Distances than those trained with gradi ent penalty-regularized and spectrally-normalized adversarial objectives. These generators also achieve nearly perfect Multi-Scale Structural Similarity Scores on CelebA, and can create high-quality samples of 128x128 images.
****

Weightless: Lossy weight encoding for deep neural network compression
Brandon Reagan, Udit Gupta, Bob Adolf, Michael Mitzenmacher, Alexander Rush, Gu-Yeon Wei, David Brooks
The large memory requirements of deep neural networks limit their deployment and adoption on many devices. Model compression methods effectively reduce the memo ry requirements of these models, usually through applying transformations such a s weight pruning or quantization. In this paper, we present a novel scheme for l ossy weight encoding co-designed with weight simplification techniques. The enco ding is based on the Bloomier filter, a probabilistic data structure that can sa ve space at the cost of introducing random errors. Leveraging the ability of neu ral networks to tolerate these imperfections and by re-training around the error s, the proposed technique, named Weightless, can compress weights by up to 496x without loss of model accuracy. This results in up to a 1.51x improvement over t he state-of-the-art.
****

Learning to Reweight Examples for Robust Deep Learning
Mengye Ren, Wenyuan Zeng, Bin Yang, Raquel Urtasun
Deep neural networks have been shown to be very powerful modeling tools for many supervised learning tasks involving complex input patterns. However, they can a lso easily overfit to training set biases and label noises. In addition to vario us regularizers, example reweighting algorithms are popular solutions to these p roblems, but they require careful tuning of additional hyperparameters, such as example mining schedules and regularization hyperparameters. In contrast to past reweighting methods, which typically consist of functions of the cost value of each example, in this work we propose a novel meta-learning algorithm that learn s to assign weights to training examples based on their gradient directions. To determine the example weights, our method performs a meta gradient descent step on the current mini-batch example weights (which are initialized from zero) to m inimize the loss on a clean unbiased validation set. Our proposed method can be easily implemented on any type of deep network, does not require any additional

hyperparameter tuning, and achieves impressive performance on class imbalance and corrupted label problems where only a small amount of clean validation data is available.
****

Learning by Playing Solving Sparse Reward Tasks from Scratch
Martin Riedmiller, Roland Hafner, Thomas Lampe, Michael Neunert, Jonas Degrave, Tom Wiele, Vlad Mnih, Nicolas Heess, Jost Tobias Springenberg
We propose Scheduled Auxiliary Control (SAC-X), a new learning paradigm in the context of Reinforcement Learning (RL). SAC-X enables learning of complex behaviors - from scratch - in the presence of multiple sparse reward signals. To this end, the agent is equipped with a set of general auxiliary tasks, that it attempts to learn simultaneously via off-policy RL. The key idea behind our method is that active (learned) scheduling and execution of auxiliary policies allows the agent to efficiently explore its environment - enabling it to excel at sparse reward RL. Our experiments in several challenging robotic manipulation settings demonstrate the power of our approach.
****

Been There, Done That: Meta-Learning with Episodic Recall
Samuel Ritter, Jane Wang, Zeb Kurth-Nelson, Siddhant Jayakumar, Charles Blundell, Razvan Pascanu, Matthew Botvinick
Meta-learning agents excel at rapidly learning new tasks from open-ended task distributions; yet, they forget what they learn about each task as soon as the next begins. When tasks reoccur {–} as they do in natural environments {–} meta-learning agents must explore again instead of immediately exploiting previously discovered solutions. We propose a formalism for generating open-ended yet repetitious environments, then develop a meta-learning architecture for solving these environments. This architecture melds the standard LSTM working memory with a differentiable neural episodic memory. We explore the capabilities of agents with this episodic LSTM in five meta-learning environments with reoccurring tasks, ranging from bandits to navigation and stochastic sequential decision problems.
****

A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music
Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, Douglas Eck
The Variational Autoencoder (VAE) has proven to be an effective model for producing semantically meaningful latent representations for natural data. However, it has thus far seen limited application to sequential data, and, as we demonstrate, existing recurrent VAE models have difficulty modeling sequences with long-term structure. To address this issue, we propose the use of a hierarchical decoder, which first outputs embeddings for subsequences of the input and then uses these embeddings to generate each subsequence independently. This structure encourages the model to utilize its latent code, thereby avoiding the "posterior collapse" problem which remains an issue for recurrent VAEs. We apply this architecture to modeling sequences of musical notes and find that it exhibits dramatically better sampling, interpolation, and reconstruction performance than a "flat" baseline model. An implementation of our "MusicVAE" is available online at https://goo.gl/magenta/musicvae-code.
****

Learning to Optimize Combinatorial Functions
Nir Rosenfeld, Eric Balkanski, Amir Globerson, Yaron Singer
Submodular functions have become a ubiquitous tool in machine learning. They are learnable from data, and can be optimized efficiently and with guarantees. Nonetheless, recent negative results show that optimizing learned surrogates of submodular functions can result in arbitrarily bad approximations of the true optimum. Our goal in this paper is to highlight the source of this hardness, and propose an alternative criterion for optimizing general combinatorial functions from sampled data. We prove a tight equivalence showing that a class of functions is optimizable if and only if it can be learned. We provide efficient and scalable optimization algorithms for several function classes of interest, and demonstrate their utility on the task of optimally choosing trending social media items.
****

Fast Information-theoretic Bayesian Optimisation
Binxin Ru, Michael A. Osborne, Mark Mcleod, Diego Granziol
Information-theoretic Bayesian optimisation techniques have demonstrated state-of-the-art performance in tackling important global optimisation problems. However, current information-theoretic approaches require many approximations in implementation, introduce often-prohibitive computational overhead and limit the choice of kernels available to model the objective. We develop a fast information-theoretic Bayesian Optimisation method, FITBO, that avoids the need for sampling the global minimiser, thus significantly reducing computational overhead. Moreover, in comparison with existing approaches, our method faces fewer constraints on kernel choice and enjoys the merits of dealing with the output space. We demonstrate empirically that FITBO inherits the performance associated with information-theoretic Bayesian optimisation, while being even faster than simpler Bayesian optimisation approaches, such as Expected Improvement.
****

Deep One-Class Classification
Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, Marius Kloft
Despite the great advances made by deep learning in many machine learning problems, there is a relative dearth of deep learning approaches for anomaly detection. Those approaches which do exist involve networks trained to perform a task other than anomaly detection, namely generative models or compression, which are in turn adapted for use in anomaly detection; they are not trained on an anomaly detection based objective. In this paper we introduce a new anomaly detection method—Deep Support Vector Data Description—, which is trained on an anomaly detection based objective. The adaptation to the deep regime necessitates that our neural network and training procedure satisfy certain properties, which we demonstrate theoretically. We show the effectiveness of our method on MNIST and CIFAR-10 image benchmark datasets as well as on the detection of adversarial examples of GTSRB stop signs.
****

Augment and Reduce: Stochastic Inference for Large Categorical Distributions
Francisco Ruiz, Michalis Titsias, Adji Bousso Dieng, David Blei
Categorical distributions are ubiquitous in machine learning, e.g., in classification, language models, and recommendation systems. However, when the number of possible outcomes is very large, using categorical distributions becomes computationally expensive, as the complexity scales linearly with the number of outcomes. To address this problem, we propose augment and reduce (A&R), a method to alleviate the computational complexity. A&R uses two ideas: latent variable augmentation and stochastic variational inference. It maximizes a lower bound on the marginal likelihood of the data. Unlike existing methods which are specific to softmax, A&R is more general and is amenable to other categorical models, such as multinomial probit. On several large-scale classification problems, we show that A&R provides a tighter bound on the marginal likelihood and has better predictive performance than existing approaches.
****

Probabilistic Boolean Tensor Decomposition
Tammo Rukat, Chris Holmes, Christopher Yau
Boolean tensor decomposition approximates data of multi-way binary relationships as product of interpretable low-rank binary factors, following the rules Boolean algebra. Here, we present its first probabilistic treatment. We facilitate scalable sampling-based posterior inference by exploitation of the combinatorial structure of the factor conditionals. Maximum a posteriori estimates consistently outperform existing non-probabilistic approaches. We show that our performance gains can partially be explained by convergence to solutions that occupy relatively large regions of the parameter space, as well as by implicit model averaging. Moreover, the Bayesian treatment facilitates model selection with much greater accuracy than the previously suggested minimum description length based approach. We investigate three real-world data sets. First, temporal interaction networks and behavioural data of university students demonstrate the inference of instr

uctive latent patterns. Next, we decompose a tensor with more than 10 Billion da
ta points, indicating relations of gene expression in cancer patients. Not only
does this demonstrate scalability, it also provides an entirely novel perspectiv
e on relational properties of continuous data and, in the present example, on th
e molecular heterogeneity of cancer. Our implementation is available on GitHub:
https://github.com/TammoR/LogicalFactorisationMachines
****

Black-Box Variational Inference for Stochastic Differential Equations
Tom Ryder, Andrew Golightly, A. Stephen McGough, Dennis Prangle
Parameter inference for stochastic differential equations is challenging due to
the presence of a latent diffusion process. Working with an Euler-Maruyama discr
etisation for the diffusion, we use variational inference to jointly learn the p
arameters and the diffusion paths. We use a standard mean-field variational appr
oximation of the parameter posterior, and introduce a recurrent neural network t
o approximate the posterior for the diffusion paths conditional on the parameter
s. This neural network learns how to provide Gaussian state transitions which br
idge between observations in a very similar way to the conditioned diffusion pro
cess. The resulting black-box inference method can be applied to any SDE system
with light tuning requirements. We illustrate the method on a Lotka-Volterra sys
tem and an epidemic model, producing accurate parameter estimates in a few hours
.
****

Spurious Local Minima are Common in Two-Layer ReLU Neural Networks
Itay Safran, Ohad Shamir
We consider the optimization problem associated with training simple ReLU neural
 networks of the form $\mathbf{x}\mapsto \sum_{i=1}^{k}\max\{0,\mathbf{w}_i^\top
 \mathbf{x}\}$ with respect to the squared loss. We provide a computer-assisted
proof that even if the input distribution is standard Gaussian, even if the dime
nsion is arbitrarily large, and even if the target values are generated by such
a network, with orthonormal parameter vectors, the problem can still have spurio
us local minima once $6\le k\le 20$. By a concentration of measure argument, thi
s implies that in high input dimensions, nearly all target networks of the relev
ant sizes lead to spurious local minima. Moreover, we conduct experiments which
show that the probability of hitting such local minima is quite high, and increa
sing with the network size. On the positive side, mild over-parameterization app
ears to drastically reduce such local minima, indicating that an over-parameteri
zation assumption is necessary to get a positive result in this setting.
****

Learning Equations for Extrapolation and Control
Subham Sahoo, Christoph Lampert, Georg Martius
We present an approach to identify concise equations from data using a shallow n
eural network approach. In contrast to ordinary black-box regression, this appro
ach allows understanding functional relations and generalizing them from observe
d data to unseen parts of the parameter space. We show how to extend the class o
f learnable equations for a recently proposed equation learning network to inclu
de divisions, and we improve the learning and model selection strategy to be use
ful for challenging real-world data. For systems governed by analytical expressi
ons, our method can in many cases identify the true underlying equation and extr
apolate to unseen domains. We demonstrate its effectiveness by experiments on a
cart-pendulum system, where only 2 random rollouts are required to learn the for
ward dynamics and successfully achieve the swing-up task.
****

Tempered Adversarial Networks
Mehdi S. M. Sajjadi, Giambattista Parascandolo, Arash Mehrjou, Bernhard Schölkop
f
Generative adversarial networks (GANs) have been shown to produce realistic samp
les from high-dimensional distributions, but training them is considered hard. A
 possible explanation for training instabilities is the inherent imbalance betwe
en the networks: While the discriminator is trained directly on both real and fa
ke samples, the generator only has control over the fake samples it produces sin

ce the real data distribution is fixed by the choice of a given dataset. We prop
ose a simple modification that gives the generator control over the real samples
 which leads to a tempered learning process for both generator and discriminator
. The real data distribution passes through a lens before being revealed to the
discriminator, balancing the generator and discriminator by gradually revealing
more detailed features necessary to produce high-quality results. The proposed m
odule automatically adjusts the learning process to the current strength of the
networks, yet is generic and easy to add to any GAN variant. In a number of expe
riments, we show that this can improve quality, stability and/or convergence spe
ed across a range of different GAN architectures (DCGAN, LSGAN, WGAN-GP).
****

Representation Tradeoffs for Hyperbolic Embeddings
Frederic Sala, Chris De Sa, Albert Gu, Christopher Re
Hyperbolic embeddings offer excellent quality with few dimensions when embedding
 hierarchical data structures. We give a combinatorial construction that embeds
trees into hyperbolic space with arbitrarily low distortion without optimization
. On WordNet, this algorithm obtains a mean-average-precision of 0.989 with only
 two dimensions, outperforming existing work by 0.11 points. We provide bounds c
haracterizing the precision-dimensionality tradeoff inherent in any hyperbolic e
mbedding. To embed general metric spaces, we propose a hyperbolic generalization
 of multidimensional scaling (h-MDS). We show how to perform exact recovery of h
yperbolic points from distances, provide a perturbation analysis, and give a rec
overy result that enables us to reduce dimensionality. Finally, we extract lesso
ns from the algorithms and theory above to design a scalable PyTorch-based imple
mentation that can handle incomplete information.
****

Graph Networks as Learnable Physics Engines for Inference and Control
Alvaro Sanchez-Gonzalez, Nicolas Heess, Jost Tobias Springenberg, Josh Merel, Ma
rtin Riedmiller, Raia Hadsell, Peter Battaglia
Understanding and interacting with everyday physical scenes requires rich knowle
dge about the structure of the world, represented either implicitly in a value o
r policy function, or explicitly in a transition model. Here we introduce a new
class of learnable models--based on graph networks--which implement an inductive b
ias for object- and relation-centric representations of complex, dynamical syste
ms. Our results show that as a forward model, our approach supports accurate pre
dictions from real and simulated data, and surprisingly strong and efficient gen
eralization, across eight distinct physical systems which we varied parametrical
ly and structurally. We also found that our inference model can perform system i
dentification. Our models are also differentiable, and support online planning v
ia gradient-based trajectory optimization, as well as offline policy optimizatio
n. Our framework offers new opportunities for harnessing and exploiting rich kno
wledge about the world, and takes a key step toward building machines with more
human-like representations of the world.
****

A Classification-Based Study of Covariate Shift in GAN Distributions
Shibani Santurkar, Ludwig Schmidt, Aleksander Madry
A basic, and still largely unanswered, question in the context of Generative Adv
ersarial Networks (GANs) is whether they are truly able to capture all the funda
mental characteristics of the distributions they are trained on. In particular,
evaluating the diversity of GAN distributions is challenging and existing method
s provide only a partial understanding of this issue. In this paper, we develop
quantitative and scalable tools for assessing the diversity of GAN distributions
. Specifically, we take a classification-based perspective and view loss of dive
rsity as a form of covariate shift introduced by GANs. We examine two specific f
orms of such shift: mode collapse and boundary distortion. In contrast to prior
work, our methods need only minimal human supervision and can be readily applied
 to state-of-the-art GANs on large, canonical datasets. Examining popular GANs u
sing our tools indicates that these GANs have significant problems in reproducin
g the more distributional properties of their training dataset.
****

TAPAS: Tricks to Accelerate (encrypted) Prediction As a Service
Amartya Sanyal, Matt Kusner, Adria Gascon, Varun Kanade
Machine learning methods are widely used for a variety of prediction problems. P
rediction as a service is a paradigm in which service providers with technologic
al expertise and computational resources may perform predictions for clients. Ho
wever, data privacy severely restricts the applicability of such services, unles
s measures to keep client data private (even from the service provider) are desi
gned. Equally important is to minimize the nature of computation and amount of c
ommunication required between client and server. Fully homomorphic encryption of
fers a way out, whereby clients may encrypt their data, and on which the server
may perform arithmetic computations. The one drawback of using fully homomorphic
 encryption is the amount of time required to evaluate large machine learning mo
dels on encrypted data. We combine several ideas from the machine learning liter
ature, particularly work on quantization and sparsification of neural networks,
together with algorithmic tools to speed-up and parallelize computation using en
crypted data.
****
Tight Regret Bounds for Bayesian Optimization in One Dimension
Jonathan Scarlett
We consider the problem of Bayesian optimization (BO) in one dimension, under a
Gaussian process prior and Gaussian sampling noise. We provide a theoretical ana
lysis showing that, under fairly mild technical assumptions on the kernel, the b
est possible cumulative regret up to time T behaves as $\Omega(\sqrt{T})$ and
$O(\sqrt{T\log T})$. This gives a tight characterization up to a $\sqrt{\log T}$
 factor, and includes the first non-trivial lower bound for noisy BO. Our assump
tions are satisfied, for example, by the squared exponential and Matérn-ν ke
rnels, with the latter requiring $\nu > 2$. Our results certify the near-optimal
ity of existing bounds (Srinivas et al., 2009) for the SE kernel, while proving
them to be strictly suboptimal for the Matérn kernel with $\nu > 2$.
****
Learning with Abandonment
Sven Schmit, Ramesh Johari
Consider a platform that wants to learn a personalized policy for each user, but
 the platform faces the risk of a user abandoning the platform if they are dissa
tisfied with the actions of the platform. For example, a platform is interested
in personalizing the number of newsletters it sends, but faces the risk that the
 user unsubscribes forever. We propose a general thresholded learning model for
scenarios like this, and discuss the structure of optimal policies. We describe
salient features of optimal personalization algorithms and how feedback the plat
form receives impacts the results. Furthermore, we investigate how the platform
can efficiently learn the heterogeneity across users by interacting with a popul
ation and provide performance guarantees.
****
Not to Cry Wolf: Distantly Supervised Multitask Learning in Critical Care
Patrick Schwab, Emanuela Keller, Carl Muroi, David J. Mack, Christian Strässle,
Walter Karlen
Patients in the intensive care unit (ICU) require constant and close supervision
. To assist clinical staff in this task, hospitals use monitoring systems that t
rigger audiovisual alarms if their algorithms indicate that a patient’s conditio
n may be worsening. However, current monitoring systems are extremely sensitive
to movement artefacts and technical errors. As a result, they typically trigger
hundreds to thousands of false alarms per patient per day - drowning the importa
nt alarms in noise and adding to the exhaustion of clinical staff. In this setti
ng, data is abundantly available, but obtaining trustworthy annotations by exper
ts is laborious and expensive. We frame the problem of false alarm reduction fro
m multivariate time series as a machine-learning task and address it with a nove
l multitask network architecture that utilises distant supervision through multi
ple related auxiliary tasks in order to reduce the number of expensive labels re
quired for training. We show that our approach leads to significant improvements
 over several state-of-the-art baselines on real-world ICU data and provide new

insights on the importance of task selection and architectural choices in distan
tly supervised multitask learning.
****
Progress & Compress: A scalable framework for continual learning
Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwins
ka, Yee Whye Teh, Razvan Pascanu, Raia Hadsell
We introduce a conceptually simple and scalable framework for continual learning
 domains where tasks are learned sequentially. Our method is constant in the num
ber of parameters and is designed to preserve performance on previously encounte
red tasks while accelerating learning progress on subsequent problems. This is a
chieved by training a network with two components: A knowledge base, capable of
solving previously encountered problems, which is connected to an active column
that is employed to efficiently learn the current task. After learning a new tas
k, the active column is distilled into the knowledge base, taking care to protec
t any previously acquired skills. This cycle of active learning (progression) fo
llowed by consolidation (compression) requires no architecture growth, no access
 to or storing of previous data or tasks, and no task-specific parameters. We de
monstrate the progress & compress approach on sequential classification of handw
ritten alphabets as well as two reinforcement learning domains: Atari games and
3D maze navigation.
****
Multi-Fidelity Black-Box Optimization with Hierarchical Partitions
Rajat Sen, Kirthevasan Kandasamy, Sanjay Shakkottai
Motivated by settings such as hyper-parameter tuning and physical simulations, w
e consider the problem of black-box optimization of a function. Multi-fidelity t
echniques have become popular for applications where exact function evaluations
are expensive, but coarse (biased) approximations are available at much lower co
st. A canonical example is that of hyper-parameter selection in a learning algor
ithm. The learning algorithm can be trained for fewer iterations – this results
in a lower cost, but its validation error is only coarsely indicative of the sam
e if the algorithm had been trained till completion. We incorporate the multi-fi
delity setup into the powerful framework of black-box optimization through hiera
rchical partitioning. We develop tree-search based multi-fidelity algorithms wit
h theoretical guarantees on simple regret. We finally demonstrate the performanc
e gains of our algorithms on both real and synthetic datasets.
****
Overcoming Catastrophic Forgetting with Hard Attention to the Task
Joan Serra, Didac Suris, Marius Miron, Alexandros Karatzoglou
Catastrophic forgetting occurs when a neural network loses the information learn
ed in a previous task after training on subsequent tasks. This problem remains a
 hurdle for artificial intelligence systems with sequential learning capabilitie
s. In this paper, we propose a task-based hard attention mechanism that preserve
s previous tasks’ information without affecting the current task’s learning. A h
ard attention mask is learned concurrently to every task, through stochastic gra
dient descent, and previous masks are exploited to condition such learning. We s
how that the proposed mechanism is effective for reducing catastrophic forgettin
g, cutting current rates by 45 to 80%. We also show that it is robust to differe
nt hyperparameter choices, and that it offers a number of monitoring capabilitie
s. The approach features the possibility to control both the stability and compa
ctness of the learned knowledge, which we believe makes it also attractive for o
nline learning or network compression applications.
****
Bounding and Counting Linear Regions of Deep Neural Networks
Thiago Serra, Christian Tjandraatmadja, Srikumar Ramalingam
We investigate the complexity of deep neural networks (DNN) that represent piece
wise linear (PWL) functions. In particular, we study the number of linear region
s, i.e. pieces, that a PWL function represented by a DNN can attain, both theore
tically and empirically. We present (i) tighter upper and lower bounds for the m
aximum number of linear regions on rectifier networks, which are exact for input
s of dimension one; (ii) a first upper bound for multi-layer maxout networks; an

d (iii) a first method to perform exact enumeration or counting of the number of regions by modeling the DNN with a mixed-integer linear formulation. These bounds come from leveraging the dimension of the space defining each linear region. The results also indicate that a deep rectifier network can only have more linear regions than every shallow counterpart with same number of neurons if that number exceeds the dimension of the input.
****

First Order Generative Adversarial Networks
Calvin Seward, Thomas Unterthiner, Urs Bergmann, Nikolay Jetchev, Sepp Hochreiter
GANs excel at learning high dimensional distributions, but they can update generator parameters in directions that do not correspond to the steepest descent direction of the objective. Prominent examples of problematic update directions include those used in both Goodfellow’s original GAN and the WGAN-GP. To formally describe an optimal update direction, we introduce a theoretical framework which allows the derivation of requirements on both the divergence and corresponding method for determining an update direction, with these requirements guaranteeing unbiased mini-batch updates in the direction of steepest descent. We propose a novel divergence which approximates the Wasserstein distance while regularizing the critic’s first order information. Together with an accompanying update direction, this divergence fulfills the requirements for unbiased steepest descent updates. We verify our method, the First Order GAN, with image generation on CelebA, LSUN and CIFAR-10 and set a new state of the art on the One Billion Word language generation task.
****

Finding Influential Training Samples for Gradient Boosted Decision Trees
Boris Sharchilev, Yury Ustinovskiy, Pavel Serdyukov, Maarten Rijke
We address the problem of finding influential training samples for a particular case of tree ensemble-based models, e.g., Random Forest (RF) or Gradient Boosted Decision Trees (GBDT). A natural way of formalizing this problem is studying how the model’s predictions change upon leave-one-out retraining, leaving out each individual training sample. Recent work has shown that, for parametric models, this analysis can be conducted in a computationally efficient way. We propose several ways of extending this framework to non-parametric GBDT ensembles under the assumption that tree structures remain fixed. Furthermore, we introduce a general scheme of obtaining further approximations to our method that balance the trade-off between performance and computational complexity. We evaluate our approaches on various experimental setups and use-case scenarios and demonstrate both the quality of our approach to finding influential training samples in comparison to the baselines and its computational efficiency.
****

Solving Partial Assignment Problems using Random Clique Complexes
Charu Sharma, Deepak Nathani, Manohar Kaul
We present an alternate formulation of the partial assignment problem as matching random clique complexes, that are higher-order analogues of random graphs, designed to provide a set of invariants that better detect higher-order structure. The proposed method creates random clique adjacency matrices for each k-skeleton of the random clique complexes and matches them, taking into account each point as the affine combination of its geometric neighborhood. We justify our solution theoretically, by analyzing the runtime and storage complexity of our algorithm along with the asymptotic behavior of the quadratic assignment problem (QAP) that is associated with the underlying random clique adjacency matrices. Experiments on both synthetic and real-world datasets, containing severe occlusions and distortions, provide insight into the accuracy, efficiency, and robustness of our approach. We outperform diverse matching algorithms by a significant margin.
****

Adafactor: Adaptive Learning Rates with Sublinear Memory Cost
Noam Shazeer, Mitchell Stern
In several recently proposed stochastic optimization methods (e.g. RMSProp, Adam, Adadelta), parameter updates are scaled by the inverse square roots of exponen

tial moving averages of squared past gradients. Maintaining these per-parameter second-moment estimators requires memory equal to the number of parameters. For the case of neural network weight matrices, we propose maintaining only the per-row and per-column sums of these moving averages, and estimating the per-parameter second moments based on these sums. We demonstrate empirically that this method produces similar results to the baseline. Secondly, we show that adaptive methods can produce larger-than-desired updates when the decay rate of the second moment accumulator is too slow. We propose update clipping and a gradually increasing decay rate scheme as remedies. Combining these methods and dropping momentum, we achieve comparable results to the published Adam regime in training the Transformer model on the WMT 2014 English-German machine translation task, while using very little auxiliary storage in the optimizer. Finally, we propose scaling the parameter updates based on the scale of the parameters themselves.
****

Locally Private Hypothesis Testing
Or Sheffet
We initiate the study of differentially private hypothesis testing in the local-model, under both the standard (symmetric) randomized-response mechanism (Warner 1965, Kasiviswanathan et al, 2008) and the newer (non-symmetric) mechanisms (Bassily & Smith, 2015, Bassily et al, 2017). First, we study the general framework of mapping each user’s type into a signal and show that the problem of finding the maximum-likelihood distribution over the signals is feasible. Then we discuss the randomized-response mechanism and show that, in essence, it maps the null- and alternative-hypotheses onto new sets, an affine translation of the original sets. We then give sample complexity bounds for identity and independence testing under randomized-response. We then move to the newer non-symmetric mechanisms and show that there too the problem of finding the maximum-likelihood distribution is feasible. Under the mechanism of Bassily et al we give identity and independence testers with better sample complexity than the testers in the symmetric case, and we also propose a χ^2-based identity tester which we investigate empirically.
****

Learning in Integer Latent Variable Models with Nested Automatic Differentiation
Daniel Sheldon, Kevin Winner, Debora Sujono
We develop nested automatic differentiation (AD) algorithms for exact inference and learning in integer latent variable models. Recently, Winner, Sujono, and Sheldon showed how to reduce marginalization in a class of integer latent variable models to evaluating a probability generating function which contains many levels of nested high-order derivatives. We contribute faster and more stable AD algorithms for this challenging problem and a novel algorithm to compute exact gradients for learning. These contributions lead to significantly faster and more accurate learning algorithms, and are the first AD algorithms whose running time is polynomial in the number of levels of nesting.
****

Towards More Efficient Stochastic Decentralized Learning: Faster Convergence and Sparse Communication
Zebang Shen, Aryan Mokhtari, Tengfei Zhou, Peilin Zhao, Hui Qian
Recently, the decentralized optimization problem is attracting growing attention. Most existing methods are deterministic with high per-iteration cost and have a convergence rate quadratically depending on the problem condition number. Besides, the dense communication is necessary to ensure the convergence even if the dataset is sparse. In this paper, we generalize the decentralized optimization problem to a monotone operator root finding problem, and propose a stochastic algorithm named DSBA that (1) converges geometrically with a rate linearly depending on the problem condition number, and (2) can be implemented using sparse communication only. Additionally, DSBA handles important learning problems like AUC-maximization which can not be tackled efficiently in the previous problem setting. Experiments on convex minimization and AUC-maximization validate the efficiency of our method.
****

An Algorithmic Framework of Variable Metric Over-Relaxed Hybrid Proximal Extra-Gradient Method
Li Shen, Peng Sun, Yitong Wang, Wei Liu, Tong Zhang
We propose a novel algorithmic framework of Variable Metric Over-Relaxed Hybrid Proximal Extra-gradient (VMOR-HPE) method with a global convergence guarantee for the maximal monotone operator inclusion problem. Its iteration complexities and local linear convergence rate are provided, which theoretically demonstrate that a large over-relaxed step-size contributes to accelerating the proposed VMOR-HPE as a byproduct. Specifically, we find that a large class of primal and primal-dual operator splitting algorithms are all special cases of VMOR-HPE. Hence, the proposed framework offers a new insight into these operator splitting algorithms. In addition, we apply VMOR-HPE to the Karush-Kuhn-Tucker (KKT) generalized equation of linear equality constrained multi-block composite convex optimization, yielding a new algorithm, namely nonsymmetric Proximal Alternating Direction Method of Multipliers with a preconditioned Extra-gradient step in which the preconditioned metric is generated by a blockwise Barzilai-Borwein line search technique (PADMM-EBB). We also establish iteration complexities of PADMM-EBB in terms of the KKT residual. Finally, we apply PADMM-EBB to handle the nonnegative dual graph regularized low-rank representation problem. Promising results on synthetic and real datasets corroborate the efficacy of PADMM-EBB.
****
A Spectral Approach to Gradient Estimation for Implicit Distributions
Jiaxin Shi, Shengyang Sun, Jun Zhu
Recently there have been increasing interests in learning and inference with implicit distributions (i.e., distributions without tractable densities). To this end, we develop a gradient estimator for implicit distributions based on Stein’s identity and a spectral decomposition of kernel operators, where the eigenfunctions are approximated by the Nystr{ö}m method. Unlike the previous works that only provide estimates at the sample points, our approach directly estimates the gradient function, thus allows for a simple and principled out-of-sample extension. We provide theoretical results on the error bound of the estimator and discuss the bias-variance tradeoff in practice. The effectiveness of our method is demonstrated by applications to gradient-free Hamiltonian Monte Carlo and variational inference with implicit distributions. Finally, we discuss the intuition behind the estimator by drawing connections between the Nystr{ö}m method and kernel PCA, which indicates that the estimator can automatically adapt to the geometry of the underlying distribution.
****
TACO: Learning Task Decomposition via Temporal Alignment for Control
Kyriacos Shiarlis, Markus Wulfmeier, Sasha Salter, Shimon Whiteson, Ingmar Posner
Many advanced Learning from Demonstration (LfD) methods consider the decomposition of complex, real-world tasks into simpler sub-tasks. By reusing the corresponding sub-policies within and between tasks, we can provide training data for each policy from different high-level tasks and compose them to perform novel ones. Existing approaches to modular LfD focus either on learning a single high-level task or depend on domain knowledge and temporal segmentation. In contrast, we propose a weakly supervised, domain-agnostic approach based on task sketches, which include only the sequence of sub-tasks performed in each demonstration. Our approach simultaneously aligns the sketches with the observed demonstrations and learns the required sub-policies. This improves generalisation in comparison to separate optimisation procedures. We evaluate the approach on multiple domains, including a simulated 3D robot arm control task using purely image-based observations. The results show that our approach performs commensurately with fully supervised approaches, while requiring significantly less annotation effort.
****
CRAFTML, an Efficient Clustering-based Random Forest for Extreme Multi-label Learning
Wissam Siblini, Pascale Kuntz, Frank Meyer
Extreme Multi-label Learning (XML) considers large sets of items described by a

number of labels that can exceed one million. Tree-based methods, which hierarch
ically partition the problem into small scale sub-problems, are particularly pro
mising in this context to reduce the learning/prediction complexity and to open
the way to parallelization. However, the current best approaches do not exploit
tree randomization which has shown its efficiency in random forests and they res
ort to complex partitioning strategies. To overcome these limits, we here introd
uce a new random forest based algorithm with a very fast partitioning approach c
alled CRAFTML. Experimental comparisons on nine datasets from the XML literature
 show that it outperforms the other tree-based approaches. Moreover with a paral
lelized implementation reduced to five cores, it is competitive with the best st
ate-of-the-art methods which run on one hundred-core machines.
****

Asynchronous Stochastic Quasi-Newton MCMC for Non-Convex Optimization
Umut Simsekli, Cagatay Yildiz, Than Huy Nguyen, Taylan Cemgil, Gael Richard
Recent studies have illustrated that stochastic gradient Markov Chain Monte Carl
o techniques have a strong potential in non-convex optimization, where local and
 global convergence guarantees can be shown under certain conditions. By buildin
g up on this recent theory, in this study, we develop an asynchronous-parallel s
tochastic L-BFGS algorithm for non-convex optimization. The proposed algorithm i
s suitable for both distributed and shared-memory settings. We provide formal th
eoretical analysis and show that the proposed method achieves an ergodic converg
ence rate of ${\cal O}(1/\sqrt{N})$ (N being the total number of iterations) a
nd it can achieve a linear speedup under certain conditions. We perform several
experiments on both synthetic and real datasets. The results support our theory
and show that the proposed algorithm provides a significant speedup over the rec
ently proposed synchronous distributed L-BFGS algorithm.
****

K-means clustering using random matrix sparsification
Kaushik Sinha
K-means clustering algorithm using Lloyd’s heuristic is one of the most commonly
 used tools in data mining and machine learning that shows promising performance
. However, it suffers from a high computational cost resulting from pairwise Euc
lidean distance computations between data points and cluster centers in each ite
ration of Lloyd’s heuristic. Main contributing factor of this computational bott
le neck is a matrix-vector multiplication step, where the matrix contains all th
e data points and the vector is a cluster center. In this paper we show that we
can randomly sparsify the original data matrix resulting in a sparse data matrix
 which can significantly speed up the above mentioned matrix vector multiplicati
on step without significantly affecting cluster quality. In particular, we show
that optimal k-means clustering solution of the sparse data matrix, obtained by
applying random matrix sparsification, results in an approximately optimal k-mea
ns clustering objective of the original data matrix. Our empirical studies on th
ree real world datasets corroborate our theoretical findings and demonstrate tha
t our proposed sparsification method can indeed achieve satisfactory clustering
performance.
****

Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotro
n
RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Sho
r, Ron Weiss, Rob Clark, Rif A. Saurous
We present an extension to the Tacotron speech synthesis architecture that learn
s a latent embedding space of prosody, derived from a reference acoustic represe
ntation containing the desired prosody. We show that conditioning Tacotron on th
is learned embedding space results in synthesized audio that matches the prosody
 of the reference signal with fine time detail even when the reference and synth
esis speakers are different. Additionally, we show that a reference prosody embe
dding can be used to synthesize text that is different from that of the referenc
e utterance. We define several quantitative and subjective metrics for evaluatin
g prosody transfer, and report results with accompanying audio samples from sing
le-speaker and 44-speaker Tacotron models on a prosody transfer task.

****
An Inference-Based Policy Gradient Method for Learning Options
Matthew Smith, Herke Hoof, Joelle Pineau
In the pursuit of increasingly intelligent learning systems, abstraction plays a vital role in enabling sophisticated decisions to be made in complex environments. The options framework provides formalism for such abstraction over sequences of decisions. However most models require that options be given a priori, presumably specified by hand, which is neither efficient, nor scalable. Indeed, it is preferable to learn options directly from interaction with the environment. Despite several efforts, this remains a difficult problem. In this work we develop a novel policy gradient method for the automatic learning of policies with options. This algorithm uses inference methods to simultaneously improve all of the options available to an agent, and thus can be employed in an off-policy manner, without observing option labels. The differentiable inference procedure employed yields options that can be easily interpreted. Empirical results confirm these attributes, and indicate that our algorithm has an improved sample efficiency relative to state-of-the-art in learning options end-to-end.
****
Accelerating Natural Gradient with Higher-Order Invariance
Yang Song, Jiaming Song, Stefano Ermon
An appealing property of the natural gradient is that it is invariant to arbitrary differentiable reparameterizations of the model. However, this invariance property requires infinitesimal steps and is lost in practical implementations with small but finite step sizes. In this paper, we study invariance properties from a combined perspective of Riemannian geometry and numerical differential equation solving. We define the order of invariance of a numerical method to be its convergence order to an invariant solution. We propose to use higher-order integrators and geodesic corrections to obtain more invariant optimization trajectories. We prove the numerical convergence properties of geodesic corrected updates and show that they can be as computational efficient as plain natural gradient. Experimentally, we demonstrate that invariance leads to faster optimization and our techniques improve on traditional natural gradient in deep neural network training and natural policy gradient for reinforcement learning.
****
Knowledge Transfer with Jacobian Matching
Suraj Srinivas, Francois Fleuret
Classical distillation methods transfer representations from a “teacher” neural network to a “student” network by matching their output activations. Recent methods also match the Jacobians, or the gradient of output activations with the input. However, this involves making some ad hoc decisions, in particular, the choice of the loss function. In this paper, we first establish an equivalence between Jacobian matching and distillation with input noise, from which we derive appropriate loss functions for Jacobian matching. We then rely on this analysis to apply Jacobian matching to transfer learning by establishing equivalence of a recent transfer learning procedure to distillation. We then show experimentally on standard image datasets that Jacobian-based penalties improve distillation, robustness to noisy inputs, and transfer learning.
****
Universal Planning Networks: Learning Generalizable Representations for Visuomotor Control
Aravind Srinivas, Allan Jabri, Pieter Abbeel, Sergey Levine, Chelsea Finn
A key challenge in complex visuomotor control is learning abstract representations that are effective for specifying goals, planning, and generalization. To this end, we introduce universal planning networks (UPN). UPNs embed differentiable planning within a goal-directed policy. This planning computation unrolls a forward model in a latent space and infers an optimal action plan through gradient descent trajectory optimization. The plan-by-gradient-descent process and its underlying representations are learned end-to-end to directly optimize a supervised imitation learning objective. We find that the representations learned are not only effective for goal-directed visual imitation via gradient-based trajectory

optimization, but can also provide a metric for specifying goals using images. The learned representations can be leveraged to specify distance-based rewards to reach new target states for model-free reinforcement learning, resulting in substantially more effective learning when solving new tasks described via image based goals. We were able to achieve successful transfer of visuomotor planning strategies across robots with significantly different morphologies and actuation capabilities. Visit https://sites.google. com/view/upn-public/home for video highlights.

****

Structured Control Nets for Deep Reinforcement Learning

Mario Srouji, Jian Zhang, Ruslan Salakhutdinov

In recent years, Deep Reinforcement Learning has made impressive advances in solving several important benchmark problems for sequential decision making. Many control applications use a generic multilayer perceptron (MLP) for non-vision parts of the policy network. In this work, we propose a new neural network architecture for the policy network representation that is simple yet effective. The proposed Structured Control Net (SCN) splits the generic MLP into two separate sub-modules: a nonlinear control module and a linear control module. Intuitively, the nonlinear control is for forward-looking and global control, while the linear control stabilizes the local dynamics around the residual of global control. We hypothesize that this will bring together the benefits of both linear and nonlinear policies: improve training sample efficiency, final episodic reward, and generalization of learned policy, while requiring a smaller network and being generally applicable to different training methods. We validated our hypothesis with competitive results on simulations from OpenAI MuJoCo, Roboschool, Atari, and a custom urban driving environment, with various ablation and generalization tests, trained with multiple black-box and policy gradient training methods. The proposed architecture has the potential to improve upon broader control tasks by incorporating problem specific priors into the architecture. As a case study, we demonstrate much improved performance for locomotion tasks by emulating the biological central pattern generators (CPGs) as the nonlinear part of the architecture.

****

Approximation Algorithms for Cascading Prediction Models

Matthew Streeter

We present an approximation algorithm that takes a pool of pre-trained models as input and produces from it a cascaded model with similar accuracy but lower average-case cost. Applied to state-of-the-art ImageNet classification models, this yields up to a 2x reduction in floating point multiplications, and up to a 6x reduction in average-case memory I/O. The auto-generated cascades exhibit intuitive properties, such as using lower-resolution input for easier images and requiring higher prediction confidence when using a computationally cheaper model.

****

Learning Low-Dimensional Temporal Representations

Bing Su, Ying Wu

Low-dimensional discriminative representations enhance machine learning methods in both performance and complexity, motivating supervised dimensionality reduction (DR) that transforms high-dimensional data to a discriminative subspace. Most DR methods require data to be i.i.d., however, in some domains, data naturally come in sequences, where the observations are temporally correlated. We propose a DR method called LT-LDA to learn low-dimensional temporal representations. We construct the separability among sequence classes by lifting the holistic temporal structures, which are established based on temporal alignments and may change in different subspaces. We jointly learn the subspace and the associated alignments by optimizing an objective which favors easily-separable temporal structures, and show that this objective is connected to the inference of alignments, thus allows an iterative solution. We provide both theoretical insight and empirical evaluation on real-world sequence datasets to show the interest of our method.

****

Exploiting the Potential of Standard Convolutional Autoencoders for Image Restor

ation by Evolutionary Search
Masanori Suganuma, Mete Ozay, Takayuki Okatani
Researchers have applied deep neural networks to image restoration tasks, in whi ch they proposed various network architectures, loss functions, and training met hods. In particular, adversarial training, which is employed in recent studies, seems to be a key ingredient to success. In this paper, we show that simple conv olutional autoencoders (CAEs) built upon only standard network components, i.e., convolutional layers and skip connections, can outperform the state-of-the-art methods which employ adversarial training and sophisticated loss functions. The secret is to search for good architectures using an evolutionary algorithm. All we did was to train the optimized CAEs by minimizing the l2 loss between reconst ructed images and their ground truths using the ADAM optimizer. Our experimental results show that this approach achieves 27.8 dB peak signal to noise ratio (PS NR) on the CelebA dataset and 33.3 dB on the SVHN dataset, compared to 22.8 dB a nd 19.0 dB provided by the former state-of-the-art methods, respectively.
****
Stagewise Safe Bayesian Optimization with Gaussian Processes
Yanan Sui, Vincent Zhuang, Joel Burdick, Yisong Yue
Enforcing safety is a key aspect of many problems pertaining to sequential decis ion making under uncertainty, which require the decisions made at every step to be both informative of the optimal decision and also safe. For example, we value both efficacy and comfort in medical therapy, and efficiency and safety in robo tic control. We consider this problem of optimizing an unknown utility function with absolute feedback or preference feedback subject to unknown safety constrai nts. We develop an efficient safe Bayesian optimization algorithm, StageOpt, tha t separates safe region expansion and utility function maximization into two dis tinct stages. Compared to existing approaches which interleave between expansion and optimization, we show that StageOpt is more efficient and naturally applica ble to a broader class of problems. We provide theoretical guarantees for both t he satisfaction of safety constraints as well as convergence to the optimal util ity value. We evaluate StageOpt on both a variety of synthetic experiments, as w ell as in clinical practice. We demonstrate that StageOpt is more effective than existing safe optimization approaches, and is able to safely and effectively op timize spinal cord stimulation therapy in our clinical experiments.
****
Neural Program Synthesis from Diverse Demonstration Videos
Shao-Hua Sun, Hyeonwoo Noh, Sriram Somasundaram, Joseph Lim
Interpreting decision making logic in demonstration videos is key to collaborati ng with and mimicking humans. To empower machines with this ability, we propose a neural program synthesizer that is able to explicitly synthesize underlying pr ograms from behaviorally diverse and visually complicated demonstration videos. We introduce a summarizer module as part of our model to improve the network’s a bility to integrate multiple demonstrations varying in behavior. We also employ a multi-task objective to encourage the model to learn meaningful intermediate r epresentations for end-to-end training. We show that our model is able to reliab ly synthesize underlying programs as well as capture diverse behaviors exhibited in demonstrations. The code is available at https://shaohua0116.github.io/demo2 program.
****
Scalable approximate Bayesian inference for particle tracking data
Ruoxi Sun, Liam Paninski
Many important datasets in physics, chemistry, and biology consist of noisy sequ ences of images of multiple moving overlapping particles. In many cases, the obs erved particles are indistinguishable, leading to unavoidable uncertainty about nearby particles’ identities. Exact Bayesian inference is intractable in this se tting, and previous approximate Bayesian methods scale poorly. Non-Bayesian appr oaches that output a single “best” estimate of the particle tracks (thus discard ing important uncertainty information) are therefore dominant in practice. Here we propose a flexible and scalable amortized approach for Bayesian inference on this task. We introduce a novel neural network method to approximate the (intrac

table) filter-backward-sample-forward algorithm for Bayesian inference in this s
etting. By varying the simulated training data for the network, we can perform i
nference on a wide variety of data types. This approach is therefore highly flex
ible and improves on the state of the art in terms of accuracy; provides uncerta
inty estimates about the particle locations and identities; and has a test run-t
ime that scales linearly as a function of the data length and number of particle
s, thus enabling Bayesian inference in arbitrarily large particle tracking datas
ets.
****

Graphical Nonconvex Optimization via an Adaptive Convex Relaxation
Qiang Sun, Kean Ming Tan, Han Liu, Tong Zhang
We consider the problem of learning high-dimensional Gaussian graphical models.
The graphical lasso is one of the most popular methods for estimating Gaussian g
raphical models. However, it does not achieve the oracle rate of convergence. In
 this paper, we propose the graphical nonconvex optimization for optimal estimat
ion in Gaussian graphical models, which is then approximated by a sequence of co
nvex programs. Our proposal is computationally tractable and produces an estimat
or that achieves the oracle rate of convergence. The statistical error introduce
d by the sequential approximation using a sequence of convex programs is clearly
 demonstrated via a contraction property. The proposed methodology is then exten
ded to modeling semiparametric graphical models. We show via numerical studies t
hat the proposed estimator outperforms other popular methods for estimating Gaus
sian graphical models.
****

Convolutional Imputation of Matrix Networks
Qingyun Sun, Mengyuan Yan, David Donoho, boyd
A matrix network is a family of matrices, with their relations modeled as a weig
hted graph. We consider the task of completing a partially observed matrix netwo
rk. The observation comes from a novel sampling scheme where a fraction of matri
ces might be completely unobserved. How can we recover the entire matrix network
 from incomplete observations? This mathematical problem arises in many applicat
ions including medical imaging and social networks. To recover the matrix networ
k, we propose a structural assumption that the matrices are low-rank after the g
raph Fourier transform on the network. We formulate a convex optimization proble
m and prove an exact recovery guarantee for the optimization problem. Furthermor
e, we numerically characterize the exact recovery regime for varying rank and sa
mpling rate and discover a new phase transition phenomenon. Then we give an iter
ative imputation algorithm to efficiently solve optimization problem and complet
e large scale matrix networks. We demonstrate the algorithm with a variety of ap
plications such as MRI and Facebook user network.
****

Differentiable Compositional Kernel Learning for Gaussian Processes
Shengyang Sun, Guodong Zhang, Chaoqi Wang, Wenyuan Zeng, Jiaman Li, Roger Grosse
The generalization properties of Gaussian processes depend heavily on the choice
 of kernel, and this choice remains a dark art. We present the Neural Kernel Net
work (NKN), a flexible family of kernels represented by a neural network. The NK
N’s architecture is based on the composition rules for kernels, so that each uni
t of the network corresponds to a valid kernel. It can compactly approximate com
positional kernel structures such as those used by the Automatic Statistician (L
loyd et al., 2014), but because the architecture is differentiable, it is end-to
-end trainable with gradient- based optimization. We show that the NKN is univer
sal for the class of stationary kernels. Empirically we demonstrate NKN’s patter
n discovery and extrapolation abilities on several tasks that depend crucially o
n identifying the underlying structure, including time series and texture extrap
olation, as well as Bayesian optimization.
****

Learning the Reward Function for a Misspecified Model
Erik Talvitie
In model-based reinforcement learning it is typical to decouple the problems of
learning the dynamics model and learning the reward function. However, when the

dynamics model is flawed, it may generate erroneous states that would never occu r in the true environment. It is not clear a priori what value the reward functi on should assign to such states. This paper presents a novel error bound that ac counts for the reward model’s behavior in states sampled from the model. This bo und is used to extend the existing Hallucinated DAgger-MC algorithm, which offer s theoretical performance guarantees in deterministic MDPs that do not assume a perfect model can be learned. Empirically, this approach to reward learning can yield dramatic improvements in control performance when the dynamics model is fl awed.

****

D^2: Decentralized Training over Decentralized Data

Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, Ji Liu

While training a machine learning model using multiple workers, each of which co llects data from its own data source, it would be useful when the data collected from different workers are unique and different. Ironically, recent analysis of decentralized parallel stochastic gradient descent (D-PSGD) relies on the assum ption that the data hosted on different workers are not too different. In this p aper, we ask the question: Can we design a decentralized parallel stochastic gra dient descent algorithm that is less sensitive to the data variance across worke rs? In this paper, we present D^2, a novel decentralized parallel stochastic g radient descent algorithm designed for large data variance \xr{among workers} (i mprecisely, “decentralized” data). The core of D^2 is a variance reduction ext ension of D-PSGD. It improves the convergence rate from $O\left({\sigma \over \s qrt{nT}} + {(n\zeta^2)^{\frac{1}{3}} \over T^{2/3}}\right)$ to $O\left({\sigma \ over \sqrt{nT}}\right)$ where ζ^{2} denotes the variance among data on dif ferent workers. As a result, D^2 is robust to data variance among workers. We empirically evaluated D^2 on image classification tasks, where each worker has access to only the data of a limited set of labels, and find that D^2 signifi cantly outperforms D-PSGD.

****

Neural Inverse Rendering for General Reflectance Photometric Stereo

Tatsunori Taniai, Takanori Maehara

We present a novel convolutional neural network architecture for photometric ste reo (Woodham, 1980), a problem of recovering 3D object surface normals from mult iple images observed under varying illuminations. Despite its long history in co mputer vision, the problem still shows fundamental challenges for surfaces with unknown general reflectance properties (BRDFs). Leveraging deep neural networks to learn complicated reflectance models is promising, but studies in this direct ion are very limited due to difficulties in acquiring accurate ground truth for training and also in designing networks invariant to permutation of input images . In order to address these challenges, we propose a physics based unsupervised learning framework where surface normals and BRDFs are predicted by the network and fed into the rendering equation to synthesize observed images. The network w eights are optimized during testing by minimizing reconstruction loss between ob served and synthesized images. Thus, our learning process does not require groun d truth normals or even pre-training on external images. Our method is shown to achieve the state-of-the-art performance on a challenging real-world scene bench mark.

****

Black Box FDR

Wesley Tansey, Yixin Wang, David Blei, Raul Rabadan

Analyzing large-scale, multi-experiment studies requires scientists to test each experimental outcome for statistical significance and then assess the results a s a whole. We present Black Box FDR (BB-FDR), an empirical-Bayes method for anal yzing multi-experiment studies when many covariates are gathered per experiment. BB-FDR learns a series of black box predictive models to boost power and contro l the false discovery rate (FDR) at two stages of study analysis. In Stage 1, it uses a deep neural network prior to report which experiments yielded significan t outcomes. In Stage 2, a separate black box model of each covariate is used to select features that have significant predictive power across all experiments. I

n benchmarks, BB-FDR outperforms competing state-of-the-art methods in both stag
es of analysis. We apply BB-FDR to two real studies on cancer drug efficacy. For
 both studies, BB-FDR increases the proportion of significant outcomes discovere
d and selects variables that reveal key genomic drivers of drug sensitivity and
resistance in cancer.
****

Best Arm Identification in Linear Bandits with Linear Dimension Dependency
Chao Tao, Saúl Blanco, Yuan Zhou
We study the best arm identification problem in linear bandits, where the mean r
eward of each arm depends linearly on an unknown d-dimensional parameter vecto
r θ, and the goal is to identify the arm with the largest expected reward
. We first design and analyze a novel randomized θ estimator based on the
 solution to the convex relaxation of an optimal G-allocation experiment desig
n problem. Using this estimator, we describe an algorithm whose sample complexit
y depends linearly on the dimension d, as well as an algorithm with sample com
plexity dependent on the reward gaps of the best d arms, matching the lower bo
und arising from the ordinary top-arm identification problem. We finally compare
 the empirical performance of our algorithms with other state-of-the-art algorit
hms in terms of both sample complexity and computational time.
****

Chi-square Generative Adversarial Network
Chenyang Tao, Liqun Chen, Ricardo Henao, Jianfeng Feng, Lawrence Carin Duke
To assess the difference between real and synthetic data, Generative Adversarial
 Networks (GANs) are trained using a distribution discrepancy measure. Three wid
ely employed measures are information-theoretic divergences, integral probabilit
y metrics, and Hilbert space discrepancy metrics. We elucidate the theoretical c
onnections between these three popular GAN training criteria and propose a novel
 procedure, called χ^2 (Chi-square) GAN, that is conceptually simple, stabl
e at training and resistant to mode collapse. Our procedure naturally generalize
s to address the problem of simultaneous matching of multiple distributions. Fur
ther, we propose a resampling strategy that significantly improves sample qualit
y, by repurposing the trained critic function via an importance weighting mechan
ism. Experiments show that the proposed procedure improves stability and converg
ence, and yields state-of-art results on a wide range of generative modeling tas
ks.
****

Lyapunov Functions for First-Order Methods: Tight Automated Convergence Guarante
es
Adrien Taylor, Bryan Van Scoy, Laurent Lessard
We present a novel way of generating Lyapunov functions for proving linear conve
rgence rates of first-order optimization methods. Our approach provably obtains
the fastest linear convergence rate that can be verified by a quadratic Lyapunov
 function (with given states), and only relies on solving a small-sized semidefi
nite program. Our approach combines the advantages of performance estimation pro
blems (PEP, due to Drori and Teboulle (2014)) and integral quadratic constraints
 (IQC, due to Lessard et al. (2016)), and relies on convex interpolation (due to
 Taylor et al. (2017c;b)).
****

Bayesian Uncertainty Estimation for Batch Normalized Deep Networks
Mattias Teye, Hossein Azizpour, Kevin Smith
We show that training a deep network using batch normalization is equivalent to
approximate inference in Bayesian models. We further demonstrate that this findi
ng allows us to make meaningful estimates of the model uncertainty using convent
ional architectures, without modifications to the network or the training proced
ure. Our approach is thoroughly validated by measuring the quality of uncertaint
y in a series of empirical experiments on different tasks. It outperforms baseli
nes with strong statistical significance, and displays competitive performance w
ith recent Bayesian approaches.
****

Decoupling Gradient-Like Learning Rules from Representations

Philip Thomas, Christoph Dann, Emma Brunskill
In machine learning, learning often corresponds to changing the parameters of a parameterized function. A learning rule is an algorithm or mathematical expression that specifies precisely how the parameters should be changed. When creating a machine learning system, we must make two decisions: what representation should be used (i.e., what parameterized function should be used) and what learning rule should be used to search through the resulting set of representable functions. In this paper we focus on gradient-like learning rules, wherein these two decisions are coupled in a subtle (and often unintentional) way. Using most learning rules, these two decisions are coupled in a subtle (and often unintentional) way. That is, using the same learning rule with two different representations that can represent the same sets of functions can result in two different outcomes. After arguing that this coupling is undesirable, particularly when using neural networks, we present a method for partially decoupling these two decisions for a broad class of gradient-like learning rules that span unsupervised learning, reinforcement learning, and supervised learning.
****

CoVeR: Learning Covariate-Specific Vector Representations with Tensor Decompositions
Kevin Tian, Teng Zhang, James Zou
Word embedding is a useful approach to capture co-occurrence structures in large text corpora. However, in addition to the text data itself, we often have additional covariates associated with individual corpus documents—e.g. the demographic of the author, time and venue of publication—and we would like the embedding to naturally capture this information. We propose CoVeR, a new tensor decomposition model for vector embeddings with covariates. CoVeR jointly learns a base embedding for all the words as well as a weighted diagonal matrix to model how each covariate affects the base embedding. To obtain author or venue-specific embedding, for example, we can then simply multiply the base embedding by the associated transformation matrix. The main advantages of our approach are data efficiency and interpretability of the covariate transformation. Our experiments demonstrate that our joint model learns substantially better covariate-specific embeddings compared to the standard approach of learning a separate embedding for each covariate using only the relevant subset of data, as well as other related methods. Furthermore, CoVeR encourages the embeddings to be “topic-aligned” in that the dimensions have specific independent meanings. This allows our covariate-specific embeddings to be compared by topic, enabling downstream differential analysis. We empirically evaluate the benefits of our algorithm on datasets, and demonstrate how it can be used to address many natural questions about covariate effects.
****

Importance Weighted Transfer of Samples in Reinforcement Learning
Andrea Tirinzoni, Andrea Sessa, Matteo Pirotta, Marcello Restelli
We consider the transfer of experience samples (i.e., tuples < s, a, s’, r >) in reinforcement learning (RL), collected from a set of source tasks to improve the learning process in a given target task. Most of the related approaches focus on selecting the most relevant source samples for solving the target task, but then all the transferred samples are used without considering anymore the discrepancies between the task models. In this paper, we propose a model-based technique that automatically estimates the relevance (importance weight) of each source sample for solving the target task. In the proposed approach, all the samples are transferred and used by a batch RL algorithm to solve the target task, but their contribution to the learning process is proportional to their importance weight. By extending the results for importance weighting provided in supervised learning literature, we develop a finite-sample analysis of the proposed batch RL algorithm. Furthermore, we empirically compare the proposed algorithm to state-of-the-art approaches, showing that it achieves better learning performance and is very robust to negative transfer, even when some source tasks are significantly different from the target task.
****

Adversarial Regression with Multiple Learners
Liang Tong, Sixie Yu, Scott Alfeld, vorobeychik
Despite the considerable success enjoyed by machine learning techniques in practice, numerous studies demonstrated that many approaches are vulnerable to attacks. An important class of such attacks involves adversaries changing features at test time to cause incorrect predictions. Previous investigations of this problem pit a single learner against an adversary. However, in many situations an adversary's decision is aimed at a collection of learners, rather than specifically targeted at each independently. We study the problem of adversarial linear regression with multiple learners. We approximate the resulting game by exhibiting an upper bound on learner loss functions, and show that the resulting game has a unique symmetric equilibrium. We present an algorithm for computing this equilibrium, and show through extensive experiments that equilibrium models are significantly more robust than conventional regularized linear regression.
****

Convergent Tree Backup and Retrace with Function Approximation
Ahmed Touati, Pierre-Luc Bacon, Doina Precup, Pascal Vincent
Off-policy learning is key to scaling up reinforcement learning as it allows to learn about a target policy from the experience generated by a different behavior policy. Unfortunately, it has been challenging to combine off-policy learning with function approximation and multi-step bootstrapping in a way that leads to both stable and efficient algorithms. In this work, we show that the Tree Backup and Retrace algorithms are unstable with linear function approximation, both in theory and in practice with specific examples. Based on our analysis, we then derive stable and efficient gradient-based algorithms using a quadratic convex-concave saddle-point formulation. By exploiting the problem structure proper to these algorithms, we are able to provide convergence guarantees and finite-sample bounds. The applicability of our new analysis also goes beyond Tree Backup and Retrace and allows us to provide new convergence rates for the GTD and GTD2 algorithms without having recourse to projections or Polyak averaging.
****

Learning Longer-term Dependencies in RNNs with Auxiliary Losses
Trieu Trinh, Andrew Dai, Thang Luong, Quoc Le
Despite recent advances in training recurrent neural networks (RNNs), capturing long-term dependencies in sequences remains a fundamental challenge. Most approaches use backpropagation through time (BPTT), which is difficult to scale to very long sequences. This paper proposes a simple method that improves the ability to capture long term dependencies in RNNs by adding an unsupervised auxiliary loss to the original objective. This auxiliary loss forces RNNs to either reconstruct previous events or predict next events in a sequence, making truncated backpropagation feasible for long sequences and also improving full BPTT. We evaluate our method on a variety of settings, including pixel-by-pixel image classification with sequence lengths up to 16000, and a real document classification benchmark. Our results highlight good performance and resource efficiency of this approach over competitive baselines, including other recurrent models and a comparable sized Transformer. Further analyses reveal beneficial effects of the auxiliary loss on optimization and regularization, as well as extreme cases where there is little to no backpropagation.
****

Theoretical Analysis of Sparse Subspace Clustering with Missing Entries
Manolis Tsakiris, Rene Vidal
Sparse Subspace Clustering (SSC) is a popular unsupervised machine learning method for clustering data lying close to an unknown union of low-dimensional linear subspaces; a problem with numerous applications in pattern recognition and computer vision. Even though the behavior of SSC for complete data is by now well-understood, little is known about its theoretical properties when applied to data with missing entries. In this paper we give theoretical guarantees for SSC with incomplete data, and provide theoretical evidence that projecting the zero-filled data onto the observation pattern of the point being expressed can lead to substantial improvement in performance; a phenomenon already known experimentally.

The main insight of our analysis is that even though this projection induces add itional missing entries, this is counterbalanced by the fact that the projected and zero-filled data are in effect incomplete points associated with the union o f the corresponding projected subspaces, with respect to which the point being e xpressed is complete. The significance of this phenomenon potentially extends to the entire class of self-expressive methods.
****

StrassenNets: Deep Learning with a Multiplication Budget
Michael Tschannen, Aran Khanna, Animashree Anandkumar
A large fraction of the arithmetic operations required to evaluate deep neural n etworks (DNNs) consists of matrix multiplications, in both convolution and fully connected layers. We perform end-to-end learning of low-cost approximations of matrix multiplications in DNN layers by casting matrix multiplications as 2-laye r sum-product networks (SPNs) (arithmetic circuits) and learning their (ternary) edge weights from data. The SPNs disentangle multiplication and addition operat ions and enable us to impose a budget on the number of multiplication operations . Combining our method with knowledge distillation and applying it to image clas sification DNNs (trained on ImageNet) and language modeling DNNs (using LSTMs), we obtain a first-of-a-kind reduction in number of multiplications (over 99.5%) while maintaining the predictive performance of the full-precision models. Final ly, we demonstrate that the proposed framework is able to rediscover Strassen’s matrix multiplication algorithm, learning to multiply 2×2 matrices usin g only 7 multiplications instead of 8.
****

Invariance of Weight Distributions in Rectified MLPs
Russell Tsuchida, Fred Roosta, Marcus Gallagher
An interesting approach to analyzing neural networks that has received renewed a ttention is to examine the equivalent kernel of the neural network. This is base d on the fact that a fully connected feedforward network with one hidden layer, a certain weight distribution, an activation function, and an infinite number of neurons can be viewed as a mapping into a Hilbert space. We derive the equivale nt kernels of MLPs with ReLU or Leaky ReLU activations for all rotationally-inva riant weight distributions, generalizing a previous result that required Gaussia n weight distributions. Additionally, the Central Limit Theorem is used to show that for certain activation functions, kernels corresponding to layers with weig ht distributions having 0 mean and finite absolute third moment are asymptotic ally universal, and are well approximated by the kernel corresponding to layers with spherical Gaussian weights. In deep networks, as depth increases the equiva lent kernel approaches a pathological fixed point, which can be used to argue wh y training randomly initialized networks can be difficult. Our results also have implications for weight initialization.
****

Least-Squares Temporal Difference Learning for the Linear Quadratic Regulator
Stephen Tu, Benjamin Recht
Reinforcement learning (RL) has been successfully used to solve many continuous control tasks. Despite its impressive results however, fundamental questions reg arding the sample complexity of RL on continuous problems remain open. We study the performance of RL in this setting by considering the behavior of the Least-S quares Temporal Difference (LSTD) estimator on the classic Linear Quadratic Regu lator (LQR) problem from optimal control. We give the first finite-time analysis of the number of samples needed to estimate the value function for a fixed stat ic state-feedback policy to within epsilon-relative error. In the process of der iving our result, we give a general characterization for when the minimum eigenv alue of the empirical covariance matrix formed along the sample path of a fast-m ixing stochastic process concentrates above zero, extending a result by Koltchin skii and Mendelson in the independent covariates setting. Finally, we provide ex perimental evidence indicating that our analysis correctly captures the qualitat ive behavior of LSTD on several LQR instances.
****

The Mirage of Action-Dependent Baselines in Reinforcement Learning

George Tucker, Surya Bhupatiraju, Shixiang Gu, Richard Turner, Zoubin Ghahramani, Sergey Levine

Policy gradient methods are a widely used class of model-free reinforcement learning algorithms where a state-dependent baseline is used to reduce gradient estimator variance. Several recent papers extend the baseline to depend on both the state and action and suggest that this significantly reduces variance and improves sample efficiency without introducing bias into the gradient estimates. To better understand this development, we decompose the variance of the policy gradient estimator and numerically show that learned state-action-dependent baselines do not in fact reduce variance over a state-dependent baseline in commonly tested benchmark domains. We confirm this unexpected result by reviewing the open-source code accompanying these prior papers, and show that subtle implementation decisions cause deviations from the methods presented in the papers and explain the source of the previously observed empirical gains. Furthermore, the variance decomposition highlights areas for improvement, which we demonstrate by illustrating a simple change to the typical value function parameterization that can significantly improve performance.

****

Adversarial Risk and the Dangers of Evaluating Against Weak Attacks

Jonathan Uesato, Brendan O'Donoghue, Pushmeet Kohli, Aaron Oord

This paper investigates recently proposed approaches for defending against adversarial examples and evaluating adversarial robustness. We motivate adversarial risk as an objective for achieving models robust to worst-case inputs. We then frame commonly used attacks and evaluation metrics as defining a tractable surrogate objective to the true adversarial risk. This suggests that models may optimize this surrogate rather than the true adversarial risk. We formalize this notion as obscurity to an adversary, and develop tools and heuristics for identifying obscured models and designing transparent models. We demonstrate that this is a significant problem in practice by repurposing gradient-free optimization techniques into adversarial attacks, which we use to decrease the accuracy of several recently proposed defenses to near zero. Our hope is that our formulations and results will help researchers to develop more powerful defenses.

****

DVAE++: Discrete Variational Autoencoders with Overlapping Transformations

Arash Vahdat, William Macready, Zhengbing Bian, Amir Khoshaman, Evgeny Andriyash

Training of discrete latent variable models remains challenging because passing gradient information through discrete units is difficult. We propose a new class of smoothing transformations based on a mixture of two overlapping distributions, and show that the proposed transformation can be used for training binary latent models with either directed or undirected priors. We derive a new variational bound to efficiently train with Boltzmann machine priors. Using this bound, we develop DVAE++, a generative model with a global discrete prior and a hierarchy of convolutional continuous variables. Experiments on several benchmarks show that overlapping transformations outperform other recent continuous relaxations of discrete latent variables including Gumbel-Softmax (Maddison et al., 2016; Jang et al., 2016), and discrete variational autoencoders (Rolfe 2016).

****

Programmatically Interpretable Reinforcement Learning

Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, Swarat Chaudhuri

We present a reinforcement learning framework, called Programmatically Interpretable Reinforcement Learning (PIRL), that is designed to generate interpretable and verifiable agent policies. Unlike the popular Deep Reinforcement Learning (DRL) paradigm, which represents policies by neural networks, PIRL represents policies using a high-level, domain-specific programming language. Such programmatic policies have the benefits of being more easily interpreted than neural networks, and being amenable to verification by symbolic methods. We propose a new method, called Neurally Directed Program Search (NDPS), for solving the challenging nonsmooth optimization problem of finding a programmatic policy with maximal reward. NDPS works by first learning a neural policy network using DRL, and then per

forming a local search over programmatic policies that seeks to minimize a distance from this neural “oracle”. We evaluate NDPS on the task of learning to drive a simulated car in the TORCS car-racing environment. We demonstrate that NDPS is able to discover human-readable policies that pass some significant performance bars. We also show that PIRL policies can have smoother trajectories, and can be more easily transferred to environments not encountered during training, than corresponding policies discovered by DRL.
****

Clustering Semi-Random Mixtures of Gaussians
Aravindan Vijayaraghavan, Pranjal Awasthi
Gaussian mixture models (GMM) are the most widely used statistical model for the k-means clustering problem and form a popular framework for clustering in machine learning and data analysis. In this paper, we propose a natural robust model for k-means clustering that generalizes the Gaussian mixture model, and that we believe will be useful in identifying robust algorithms. Our first contribution is a polynomial time algorithm that provably recovers the ground-truth up to small classification error w.h.p., assuming certain separation between the components. Perhaps surprisingly, the algorithm we analyze is the popular Lloyd’s algorithm for k-means clustering that is the method-of-choice in practice. Our second result complements the upper bound by giving a nearly matching lower bound on the number of misclassified points incurred by any k-means clustering algorithm on the semi-random model.
****

A Probabilistic Theory of Supervised Similarity Learning for Pointwise ROC Curve Optimization
Robin Vogel, Aurélien Bellet, Stéphan Clémençon
The performance of many machine learning techniques depends on the choice of an appropriate similarity or distance measure on the input space. Similarity learning (or metric learning) aims at building such a measure from training data so that observations with the same (resp. different) label are as close (resp. far) as possible. In this paper, similarity learning is investigated from the perspective of pairwise bipartite ranking, where the goal is to rank the elements of a database by decreasing order of the probability that they share the same label with some query data point, based on the similarity scores. A natural performance criterion in this setting is pointwise ROC optimization: maximize the true positive rate under a fixed false positive rate. We study this novel perspective on similarity learning through a rigorous probabilistic framework. The empirical version of the problem gives rise to a constrained optimization formulation involving U-statistics, for which we derive universal learning rates as well as faster rates under a noise assumption on the data distribution. We also address the large-scale setting by analyzing the effect of sampling-based approximations. Our theoretical results are supported by illustrative numerical experiments.
****

Hierarchical Multi-Label Classification Networks
Jonatas Wehrmann, Ricardo Cerri, Rodrigo Barros
One of the most challenging machine learning problems is a particular case of data classification in which classes are hierarchically structured and objects can be assigned to multiple paths of the class hierarchy at the same time. This task is known as hierarchical multi-label classification (HMC), with applications in text classification, image annotation, and in bioinformatics problems such as protein function prediction. In this paper, we propose novel neural network architectures for HMC called HMCN, capable of simultaneously optimizing local and global loss functions for discovering local hierarchical class-relationships and global information from the entire class hierarchy while penalizing hierarchical violations. We evaluate its performance in 21 datasets from four distinct domains, and we compare it against the current HMC state-of-the-art approaches. Results show that HMCN substantially outperforms all baselines with statistical significance, arising as the novel state-of-the-art for HMC.
****

Transfer Learning via Learning to Transfer

Ying WEI, Yu Zhang, Junzhou Huang, Qiang Yang
In transfer learning, what and how to transfer are two primary issues to be addr
essed, as different transfer learning algorithms applied between a source and a
target domain result in different knowledge transferred and thereby the performa
nce improvement in the target domain. Determining the optimal one that maximizes
the performance improvement requires either exhaustive exploration or considera
ble expertise. Meanwhile, it is widely accepted in educational psychology that h
uman beings improve transfer learning skills of deciding what to transfer throug
h meta-cognitive reflection on inductive transfer learning practices. Motivated
by this, we propose a novel transfer learning framework known as Learning to Tra
nsfer (L2T) to automatically determine what and how to transfer are the best by
leveraging previous transfer learning experiences. We establish the L2T framewor
k in two stages: 1) we learn a reflection function encrypting transfer learning
skills from experiences; and 2) we infer what and how to transfer are the best f
or a future pair of domains by optimizing the reflection function. We also theor
etically analyse the algorithmic stability and generalization bound of L2T, and
empirically demonstrate its superiority over several state-of-the-art transfer l
earning algorithms.
****

Semi-Supervised Learning on Data Streams via Temporal Label Propagation
Tal Wagner, Sudipto Guha, Shiva Kasiviswanathan, Nina Mishra
We consider the problem of labeling points on a fast-moving data stream when onl
y a small number of labeled examples are available. In our setting, incoming poi
nts must be processed efficiently and the stream is too large to store in its en
tirety. We present a semi-supervised learning algorithm for this task. The algor
ithm maintains a small synopsis of the stream which can be quickly updated as ne
w points arrive, and labels every incoming point by provably learning from the f
ull history of the stream. Experiments on real datasets validate that the algori
thm can quickly and accurately classify points on a stream with a small quantity
of labeled examples.
****

Neural Dynamic Programming for Musical Self Similarity
Christian Walder, Dongwoo Kim
We present a neural sequence model designed specifically for symbolic music. The
model is based on a learned edit distance mechanism which generalises a classic
recursion from computer science, leading to a neural dynamic program. Repeated
motifs are detected by learning the transformations between them. We represent t
he arising computational dependencies using a novel data structure, the edit tre
e; this perspective suggests natural approximations which afford the scaling up
of our otherwise cubic time algorithm. We demonstrate our model on real and synt
hetic data; in all cases it out-performs a strong stacked long short-term memory
benchmark.
****

Thompson Sampling for Combinatorial Semi-Bandits
Siwei Wang, Wei Chen
We study the application of the Thompson sampling (TS) methodology to the stocha
stic combinatorial multi-armed bandit (CMAB) framework. We analyze the standard
TS algorithm for the general CMAB, and obtain the first distribution-dependent r
egret bound of $O(m\log T / \Delta_{\min}) $ for TS under general CMAB, where $m
$ is the number of arms, T is the time horizon, and $\Delta_{\min}$ is the min
imum gap between the expected reward of the optimal solution and any non-optimal
solution. We also show that one cannot use an approximate oracle in TS algorith
m for even MAB problems. Then we expand the analysis to matroid bandit, a specia
l case of CMAB and for which we could remove the independence assumption across
arms and achieve a better regret bound. Finally, we use some experiments to show
the comparison of regrets of CUCB and CTS algorithms.
****

PredRNN++: Towards A Resolution of the Deep-in-Time Dilemma in Spatiotemporal Pr
edictive Learning
Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, Philip S Yu

We present PredRNN++, a recurrent network for spatiotemporal predictive learning . In pursuit of a great modeling capability for short-term video dynamics, we ma ke our network deeper in time by leveraging a new recurrent structure named Caus al LSTM with cascaded dual memories. To alleviate the gradient propagation diffi culties in deep predictive models, we propose a Gradient Highway Unit, which pro vides alternative quick routes for the gradient flows from outputs back to long- range previous inputs. The gradient highway units work seamlessly with the causa l LSTMs, enabling our model to capture the short-term and the long-term video de pendencies adaptively. Our model achieves state-of-the-art prediction results on both synthetic and real video datasets, showing its power in modeling entangled motions.

****

Analyzing the Robustness of Nearest Neighbors to Adversarial Examples

Yizhen Wang, Somesh Jha, Kamalika Chaudhuri

Motivated by safety-critical applications, test-time attacks on classifiers via adversarial examples has recently received a great deal of attention. However, t here is a general lack of understanding on why adversarial examples arise; wheth er they originate due to inherent properties of data or due to lack of training samples remains ill-understood. In this work, we introduce a theoretical framewo rk analogous to bias-variance theory for understanding these effects. We use our framework to analyze the robustness of a canonical non-parametric classifier {– } the k-nearest neighbors. Our analysis shows that its robustness properties dep end critically on the value of k {–} the classifier may be inherently non-robust for small k, but its robustness approaches that of the Bayes Optimal classifier for fast-growing k. We propose a novel modified 1-nearest neighbor classifier, and guarantee its robustness in the large sample limit. Our experiments suggest that this classifier may have good robustness properties even for reasonable dat a set sizes.

****

Competitive Multi-agent Inverse Reinforcement Learning with Sub-optimal Demonstr ations

Xingyu Wang, Diego Klabjan

This paper considers the problem of inverse reinforcement learning in zero-sum s tochastic games when expert demonstrations are known to be suboptimal. Compared to previous works that decouple agents in the game by assuming optimality in exp ert policies, we introduce a new objective function that directly pits experts a gainst Nash Equilibrium policies, and we design an algorithm to solve for the re ward function in the context of inverse reinforcement learning with deep neural networks as model approximations. To ?nd Nash Equilibrium in large-scale games, we also propose an adversarial training algorithm for zero-sum stochastic games, and show the theoretical appeal of non-existence of local optima in its objecti ve function. In numerical experiments, we demonstrate that our Nash Equilibrium and inverse reinforcement learning algorithms address games that are not amenabl e to existing benchmark algorithms. Moreover, our algorithm successfully recover s reward and policy functions regardless of the quality of the sub-optimal exper t demonstration set.

****

Coded Sparse Matrix Multiplication

Sinong Wang, Jiashang Liu, Ness Shroff

In a large-scale and distributed matrix multiplication problem $C=A^{\intercal}B $, where $C\in\mathbb{R}^{r\times t}$, the coded computation plays an important role to effectively deal with “stragglers” (distributed computations that may ge t delayed due to few slow or faulty processors). However, existing coded schemes could destroy the significant sparsity that exists in large-scale machine learn ing problems, and could result in much higher computation overhead, i.e., $O(rt) $ decoding time. In this paper, we develop a new coded computation strategy, we call sparse code, which achieves near optimal recovery threshold, low computatio n overhead, and linear decoding time $O(nnz(C))$. We implement our scheme and de monstrate the advantage of the approach over both uncoded and current fastest co ded strategies.

****
A Fast and Scalable Joint Estimator for Integrating Additional Knowledge in Lear ning Multiple Related Sparse Gaussian Graphical Models
Beilun Wang, Arshdeep Sekhon, Yanjun Qi
We consider the problem of including additional knowledge in estimating sparse G aussian graphical models (sGGMs) from aggregated samples, arising often in bioin formatics and neuroimaging applications. Previous joint sGGM estimators either f ail to use existing knowledge or cannot scale-up to many tasks (large K) under a high-dimensional (large p) situation. In this paper, we propose a novel \un derline{J}oint \underline{E}lementary \underline{E}stimator incorporating additi onal \underline{K}nowledge (JEEK) to infer multiple related sparse Gaussian Grap hical models from large-scale heterogeneous data. Using domain knowledge as weig hts, we design a novel hybrid norm as the minimization objective to enforce the superposition of two weighted sparsity constraints, one on the shared interactio ns and the other on the task-specific structural patterns. This enables JEEK to elegantly consider various forms of existing knowledge based on the domain at ha nd and avoid the need to design knowledge-specific optimization. JEEK is solved through a fast and entry-wise parallelizable solution that largely improves the computational efficiency of the state-of-the-art $O(p^5K^4)$ to $O(p^2K^4)$. We conduct a rigorous statistical analysis showing that JEEK achieves the same conv ergence rate $O(\log(Kp)/n_{tot})$ as the state-of-the-art estimators that are m uch harder to compute. Empirically, on multiple synthetic datasets and one real- world data from neuroscience, JEEP outperforms the speed of the state-of-arts si gnificantly while achieving the same level of prediction accuracy.
****
Provable Variable Selection for Streaming Features
Jing Wang, Jie Shen, Ping Li
In large-scale machine learning applications and high-dimensional statistics, it is ubiquitous to address a considerable number of features among which many are redundant. As a remedy, online feature selection has attracted increasing atten tion in recent years. It sequentially reveals features and evaluates the importa nce of them. Though online feature selection has proven an elegant methodology, it is usually challenging to carry out a rigorous theoretical characterization. In this work, we propose a provable online feature selection algorithm that util izes the online leverage score. The selected features are then fed to k-means clustering, making the clustering step memory and computationally efficient. We prove that with high probability, performing k-means clustering based on the s elected feature space does not deviate far from the optimal clustering using the original data. The empirical results on real-world data sets demonstrate the ef fectiveness of our algorithm.
****
Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Sp eech Synthesis
Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor , Ying Xiao, Ye Jia, Fei Ren, Rif A. Saurous
In this work, we propose “global style tokens” (GSTs), a bank of embeddings that are jointly trained within Tacotron, a state-of-the-art end-to-end speech synth esis system. The embeddings are trained with no explicit labels, yet learn to mo del a large range of acoustic expressiveness. GSTs lead to a rich set of signifi cant results. The soft interpretable “labels” they generate can be used to contr ol synthesis in novel ways, such as varying speed and speaking style – independe ntly of the text content. They can also be used for style transfer, replicating the speaking style of a single audio clip across an entire long-form text corpus . When trained on noisy, unlabeled found data, GSTs learn to factorize noise and speaker identity, providing a path towards highly scalable but robust speech sy nthesis.
****
Adversarial Distillation of Bayesian Neural Network Posteriors
Kuan-Chieh Wang, Paul Vicol, James Lucas, Li Gu, Roger Grosse, Richard Zemel
Bayesian neural networks (BNNs) allow us to reason about uncertainty in a princi

pled way. Stochastic Gradient Langevin Dynamics (SGLD) enables efficient BNN lea rning by drawing samples from the BNN posterior using mini-batches. However, SGL D and its extensions require storage of many copies of the model parameters, a p otentially prohibitive cost, especially for large neural networks. We propose a framework, Adversarial Posterior Distillation, to distill the SGLD samples using a Generative Adversarial Network (GAN). At test-time, samples are generated by the GAN. We show that this distillation framework incurs no loss in performance on recent BNN applications including anomaly detection, active learning, and def ense against adversarial attacks. By construction, our framework distills not on ly the Bayesian predictive distribution, but the posterior itself. This allows o ne to compute quantities such as the approximate model variance, which is useful in downstream tasks. To our knowledge, these are the first results applying MCM C-based BNNs to the aforementioned applications.

****

Minimax Concave Penalized Multi-Armed Bandit Model with High-Dimensional Covaria tes

Xue Wang, Mingcheng Wei, Tao Yao

In this paper, we propose a Minimax Concave Penalized Multi-Armed Bandit (MCP-Ba ndit) algorithm for a decision-maker facing high-dimensional data with latent sp arse structure in an online learning and decision-making process. We demonstrate that the MCP-Bandit algorithm asymptotically achieves the optimal cumulative re gret in sample size T, O(log T), and further attains a tighter bound in both cov ariates dimension d and the number of significant covariates s, O(s^2 (s + log d). In addition, we develop a linear approximation method, the 2-step Weighted La sso procedure, to identify the MCP estimator for the MCP-Bandit algorithm under non-i.i.d. samples. Using this procedure, the MCP estimator matches the oracle e stimator with high probability. Finally, we present two experiments to benchmark our proposed the MCP-Bandit algorithm to other bandit algorithms. Both experime nts demonstrate that the MCP-Bandit algorithm performs favorably over other benc hmark algorithms, especially when there is a high level of data sparsity or when the sample size is not too small.

****

Online Convolutional Sparse Coding with Sample-Dependent Dictionary

Yaqing Wang, Quanming Yao, James Tin-Yau Kwok, Lionel M. NI

Convolutional sparse coding (CSC) has been popularly used for the learning of sh ift-invariant dictionaries in image and signal processing. However, existing met hods have limited scalability. In this paper, instead of convolving with a dicti onary shared by all samples, we propose the use of a sample-dependent dictionary in which each filter is a linear combination of a small set of base filters lea rned from data. This added flexibility allows a large number of sample-dependent patterns to be captured, which is especially useful in the handling of large or high-dimensional data sets. Computationally, the resultant model can be efficie ntly learned by online learning. Extensive experimental results on a number of d ata sets show that the proposed method outperforms existing CSC algorithms with significantly reduced time and space complexities.

****

Stein Variational Message Passing for Continuous Graphical Models

Dilin Wang, Zhe Zeng, Qiang Liu

We propose a novel distributed inference algorithm for continuous graphical mode ls, by extending Stein variational gradient descent (SVGD) to leverage the Marko v dependency structure of the distribution of interest. Our approach combines SV GD with a set of structured local kernel functions defined on the Markov blanket of each node, which alleviates the curse of high dimensionality and simultaneou sly yields a distributed algorithm for decentralized inference tasks. We justify our method with theoretical analysis and show that the use of local kernels can be viewed as a new type of localized approximation that matches the target dist ribution on the conditional distributions of each node over its Markov blanket. Our empirical results show that our method outperforms a variety of baselines in cluding standard MCMC and particle message passing methods.

****

Approximate Leave-One-Out for Fast Parameter Tuning in High Dimensions
Shuaiwen Wang, Wenda Zhou, Haihao Lu, Arian Maleki, Vahab Mirrokni
We study the parameter tuning problem for the penalized regression model. Findin g the optimal choice of the regularization parameter is a challenging problem in high-dimensional regimes where both the number of observations n and the number of parameters p are large. We propose two frameworks to obtain a computationall y efficient approximation ALO of the leave-one-out cross validation (LOOCV) risk for nonsmooth losses and regularizers. Our two frameworks are based on the prim al and dual formulations of the penalized regression model. We prove the equival ence of the two approaches under smoothness conditions. This equivalence enables us to justify the accuracy of both methods under such conditions. We use our ap proaches to obtain a risk estimate for several standard problems, including gene ralized LASSO, nuclear norm regularization and support vector machines. We exper imentally demonstrate the effectiveness of our results for non-differentiable ca ses.
****
Curriculum Learning by Transfer Learning: Theory and Experiments with Deep Netwo rks
Daphna Weinshall, Gad Cohen, Dan Amir
We provide theoretical investigation of curriculum learning in the context of st ochastic gradient descent when optimizing the convex linear regression loss. We prove that the rate of convergence of an ideal curriculum learning method is mon otonically increasing with the difficulty of the examples. Moreover, among all e qually difficult points, convergence is faster when using points which incur hig her loss with respect to the current hypothesis. We then analyze curriculum lear ning in the context of training a CNN. We describe a method which infers the cur riculum by way of transfer learning from another network, pre-trained on a diffe rent task. While this approach can only approximate the ideal curriculum, we obs erve empirically similar behavior to the one predicted by the theory, namely, a significant boost in convergence speed at the beginning of training. When the ta sk is made more difficult, improvement in generalization performance is also obs erved. Finally, curriculum learning exhibits robustness against unfavorable cond itions such as excessive regularization.
****
Extracting Automata from Recurrent Neural Networks Using Queries and Counterexam ples
Gail Weiss, Yoav Goldberg, Eran Yahav
We present a novel algorithm that uses exact learning and abstraction to extract a deterministic finite automaton describing the state dynamics of a given train ed RNN. We do this using Angluin’s \lstar algorithm as a learner and the trained RNN as an oracle. Our technique efficiently extracts accurate automata from tra ined RNNs, even when the state vectors are large and require fine differentiatio n.
****
LeapsAndBounds: A Method for Approximately Optimal Algorithm Configuration
Gellert Weisz, Andras Gyorgy, Csaba Szepesvari
We consider the problem of configuring general-purpose solvers to run efficientl y on problem instances drawn from an unknown distribution. The goal of the confi gurator is to find a configuration that runs fast on average on most instances, and do so with the least amount of total work. It can run a chosen solver on a r andom instance until the solver finishes or a timeout is reached. We propose Lea psAndBounds, an algorithm that tests configurations on randomly selected problem instances for longer and longer time. We prove that the capped expected runtime of the configuration returned by LeapsAndBounds is close to the optimal expecte d runtime, while our algorithm’s running time is near-optimal. Our results show that LeapsAndBounds is more efficient than the recent algorithm of Kleinberg et al. (2017), which, to our knowledge, is the only other algorithm configuration m ethod with non-trivial theoretical guarantees. Experimental results on configuri ng a public SAT solver on a new benchmark dataset also stand witness to the supe riority of our method.

****
Deep Predictive Coding Network for Object Recognition
Haiguang Wen, Kuan Han, Junxing Shi, Yizhen Zhang, Eugenio Culurciello, Zhongming Liu
Based on the predictive coding theory in neuro- science, we designed a bi-directional and recur- rent neural net, namely deep predictive coding networks (PCN), that has feedforward, feedback, and recurrent connections. Feedback connections from a higher layer carry the prediction of its lower-layer representation; feedforward connec- tions carry the prediction errors to its higher-layer. Given image input, PCN runs recursive cycles of bottom-up and top-down computation to update its internal representations and reduce the differ- ence between bottom-up input and top-down pre- diction at every layer. After multiple cycles of recursive updating, the representation is used for image classification. With benchmark datasets (CIFAR-10/100, SVHN, and MNIST), PCN was found to always outperform its feedforward-only counterpart: a model without any mechanism for recurrent dynamics, and its performance tended to improve given more cycles of computation over time. In short, PCN reuses a single architecture to recursively run bottom-up and top-down pro- cesses to refine its representation towards more accurate and definitive object recognition.
****
Towards Fast Computation of Certified Robustness for ReLU Networks
Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, Inderjit Dhillon
Verifying the robustness property of a general Rectified Linear Unit (ReLU) network is an NP-complete problem. Although finding the exact minimum adversarial distortion is hard, giving a certified lower bound of the minimum distortion is possible. Current available methods of computing such a bound are either time-consuming or deliver low quality bounds that are too loose to be useful. In this paper, we exploit the special structure of ReLU networks and provide two computationally efficient algorithms (Fast-Lin, Fast-Lip) that are able to certify non-trivial lower bounds of minimum adversarial distortions. Experiments show that (1) our methods deliver bounds close to (the gap is 2-3X) exact minimum distortions found by Reluplex in small networks while our algorithms are more than 10,000 times faster; (2) our methods deliver similar quality of bounds (the gap is within 35% and usually around 10%; sometimes our bounds are even better) for larger networks compared to the methods based on solving linear programming problems but our algorithms are 33-14,000 times faster; (3) our method is capable of solving large MNIST and CIFAR networks up to 7 layers with more than 10,000 neurons within tens of seconds on a single CPU core. In addition, we show that there is no polynomial time algorithm that can approximately find the minimum ℓ_1 adversarial distortion of a ReLU network with a $0.99\ln n$ approximation ratio unless NP=P, where n is the number of neurons in the network.
****
Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope
Eric Wong, Zico Kolter
We propose a method to learn deep ReLU-based classifiers that are provably robust against norm-bounded adversarial perturbations on the training data. For previously unseen examples, the approach is guaranteed to detect all adversarial examples, though it may flag some non-adversarial examples as well. The basic idea is to consider a convex outer approximation of the set of activations reachable through a norm-bounded perturbation, and we develop a robust optimization procedure that minimizes the worst case loss over this outer region (via a linear program). Crucially, we show that the dual problem to this linear program can be represented itself as a deep network similar to the backpropagation network, leading to very efficient optimization approaches that produce guaranteed bounds on the robust loss. The end result is that by executing a few more forward and backward passes through a slightly modified version of the original network (though possibly with much larger batch sizes), we can learn a classifier that is provably robust to any norm-bounded adversarial attack. We illustrate the approach on a n

umber of tasks to train classifiers with robust adversarial guarantees (e.g. for MNIST, we produce a convolutional classifier that provably has less than 5.8% test error for any adversarial attack with bounded ℓ_∞ norm less than $\epsilon = 0.1$).

****

Local Density Estimation in High Dimensions

Xian Wu, Moses Charikar, Vishnu Natchu

An important question that arises in the study of high dimensional vector representations learned from data is: given a set D of vectors and a query q, estimate the number of points within a specified distance threshold of q. Our algorithm uses locality sensitive hashing to preprocess the data to accurately and efficiently estimate the answers to such questions via an unbiased estimator that uses importance sampling. A key innovation is the ability to maintain a small number of hash tables via preprocessing data structures and algorithms that sample from multiple buckets in each hash table. We give bounds on the space requirements and query complexity of our scheme, and demonstrate the effectiveness of our algorithm by experiments on a standard word embedding dataset.

****

Adaptive Exploration-Exploitation Tradeoff for Opportunistic Bandits

Huasen Wu, Xueying Guo, Xin Liu

In this paper, we propose and study opportunistic bandits - a new variant of bandits where the regret of pulling a suboptimal arm varies under different environmental conditions, such as network load or produce price. When the load/price is low, so is the cost/regret of pulling a suboptimal arm (e.g., trying a suboptimal network configuration). Therefore, intuitively, we could explore more when the load/price is low and exploit more when the load/price is high. Inspired by this intuition, we propose an Adaptive Upper-Confidence-Bound (AdaUCB) algorithm to adaptively balance the exploration-exploitation tradeoff for opportunistic bandits. We prove that AdaUCB achieves O(log T) regret with a smaller coefficient than the traditional UCB algorithm. Furthermore, AdaUCB achieves O(1) regret with respect to T if the exploration cost is zero when the load level is below a certain threshold. Last, based on both synthetic data and real-world traces, experimental results show that AdaUCB significantly outperforms other bandit algorithms, such as UCB and TS (Thompson Sampling), under large load/price fluctuations.

****

SQL-Rank: A Listwise Approach to Collaborative Ranking

Liwei Wu, Cho-Jui Hsieh, James Sharpnack

In this paper, we propose a listwise approach for constructing user-specific rankings in recommendation systems in a collaborative fashion. We contrast the listwise approach to previous pointwise and pairwise approaches, which are based on treating either each rating or each pairwise comparison as an independent instance respectively. By extending the work of ListNet (Cao et al., 2007), we cast listwise collaborative ranking as maximum likelihood under a permutation model which applies probability mass to permutations based on a low rank latent score matrix. We present a novel algorithm called SQL-Rank, which can accommodate ties and missing data and can run in linear time. We develop a theoretical framework for analyzing listwise ranking methods based on a novel representation theory for the permutation model. Applying this framework to collaborative ranking, we derive asymptotic statistical rates as the number of users and items grow together. We conclude by demonstrating that our SQL-Rank method often outperforms current state-of-the-art algorithms for implicit feedback such as Weighted-MF and BPR and achieve favorable results when compared to explicit feedback algorithms such as matrix factorization and collaborative ranking.

****

Error Compensated Quantized SGD and its Applications to Large-scale Distributed Optimization

Jiaxiang Wu, Weidong Huang, Junzhou Huang, Tong Zhang

Large-scale distributed optimization is of great importance in various applications. For data-parallel based distributed learning, the inter-node gradient communication often becomes the performance bottleneck. In this paper, we propose the

error compensated quantized stochastic gradient descent algorithm to improve the training efficiency. Local gradients are quantized to reduce the communication overhead, and accumulated quantization error is utilized to speed up the convergence. Furthermore, we present theoretical analysis on the convergence behaviour, and demonstrate its advantage over competitors. Extensive experiments indicate that our algorithm can compress gradients by a factor of up to two magnitudes without performance degradation.
****
Reinforcing Adversarial Robustness using Model Confidence Induced by Adversarial Training
Xi Wu, Uyeong Jang, Jiefeng Chen, Lingjiao Chen, Somesh Jha
In this paper we study leveraging confidence information induced by adversarial training to reinforce adversarial robustness of a given adversarially trained model. A natural measure of confidence is $\|F(x)\|_\infty$ (i.e. how confident F is about its prediction?). We start by analyzing an adversarial training formulation proposed by Madry et al.. We demonstrate that, under a variety of instantiations, an only somewhat good solution to their objective induces confidence to be a discriminator, which can distinguish between right and wrong model predictions in a neighborhood of a point sampled from the underlying distribution. Based on this, we propose Highly Confident Near Neighbor (HCNN) a framework that combines confidence information and nearest neighbor search, to reinforce adversarial robustness of a base model. We give algorithms in this framework and perform a detailed empirical study. We report encouraging experimental results that support our analysis, and also discuss problems we observed with existing adversarial training.
****
Discrete-Continuous Mixtures in Probabilistic Programming: Generalized Semantics and Inference Algorithms
Yi Wu, Siddharth Srivastava, Nicholas Hay, Simon Du, Stuart Russell
Despite the recent successes of probabilistic programming languages (PPLs) in AI applications, PPLs offer only limited support for random variables whose distributions combine discrete and continuous elements. We develop the notion of measure-theoretic Bayesian networks (MTBNs) and use it to provide more general semantics for PPLs with arbitrarily many random variables defined over arbitrary measure spaces. We develop two new general sampling algorithms that are provably correct under the MTBN framework: the lexicographic likelihood weighting (LLW) for general MTBNs and the lexicographic particle filter (LPF), a specialized algorithm for state-space models. We further integrate MTBNs into a widely used PPL system, BLOG, and verify the effectiveness of the new inference algorithms through representative examples.
****
Variance Regularized Counterfactual Risk Minimization via Variational Divergence Minimization
Hang Wu, May Wang
Off-policy learning, the task of evaluating and improving policies using historic data collected from a logging policy, is important because on-policy evaluation is usually expensive and has adverse impacts. One of the major challenge of off-policy learning is to derive counterfactual estimators that also has low variance and thus low generalization error. In this work, inspired by learning bounds for importance sampling problems, we present a new counterfactual learning principle for off-policy learning with bandit feedbacks. Our method regularizes the generalization error by minimizing the distribution divergence between the logging policy and the new policy, and removes the need for iterating through all training samples to compute sample variance regularization in prior work. With neural network policies, our end-to-end training algorithms using variational divergence minimization showed significant improvement over conventional baseline algorithms and is also consistent with our theoretical results.
****
Deep k-Means: Re-Training and Parameter Sharing with Harder Cluster Assignments for Compressing Deep Convolutions

Junru Wu, Yue Wang, Zhenyu Wu, Zhangyang Wang, Ashok Veeraraghavan, Yingyan Lin
The current trend of pushing CNNs deeper with convolutions has created a pressing demand to achieve higher compression gains on CNNs where convolutions dominate the computation and parameter amount (e.g., GoogLeNet, ResNet and Wide ResNet). Further, the high energy consumption of convolutions limits its deployment on mobile devices. To this end, we proposed a simple yet effective scheme for compressing convolutions though applying k-means clustering on the weights, compression is achieved through weight-sharing, by only recording K cluster centers and weight assignment indexes. We then introduced a novel spectrally relaxed k-means regularization, which tends to make hard assignments of convolutional layer weights to K learned cluster centers during re-training. We additionally propose an improved set of metrics to estimate energy consumption of CNN hardware implementations, whose estimation results are verified to be consistent with previously proposed energy estimation tool extrapolated from actual hardware measurements. We finally evaluated Deep k-Means across several CNN models in terms of both compression ratio and energy consumption reduction, observing promising results without incurring accuracy loss. The code is available at https://github.com/Sandbox3aster/Deep-K-Means
****
Bayesian Quadrature for Multiple Related Integrals
Xiaoyue Xi, Francois-Xavier Briol, Mark Girolami
Bayesian probabilistic numerical methods are a set of tools providing posterior distributions on the output of numerical methods. The use of these methods is usually motivated by the fact that they can represent our uncertainty due to incomplete/finite information about the continuous mathematical problem being approximated. In this paper, we demonstrate that this paradigm can provide additional advantages, such as the possibility of transferring information between several numerical methods. This allows users to represent uncertainty in a more faithful manner and, as a by-product, provide increased numerical efficiency. We propose the first such numerical method by extending the well-known Bayesian quadrature algorithm to the case where we are interested in computing the integral of several related functions. We then prove convergence rates for the method in the well-specified and misspecified cases, and demonstrate its efficiency in the context of multi-fidelity models for complex engineering systems and a problem of global illumination in computer graphics.
****
Model-Level Dual Learning
Yingce Xia, Xu Tan, Fei Tian, Tao Qin, Nenghai Yu, Tie-Yan Liu
Many artificial intelligence tasks appear in dual forms like English$\leftrightarrow$French translation and speech$\leftrightarrow$text transformation. Existing dual learning schemes, which are proposed to solve a pair of such dual tasks, explore how to leverage such dualities from data level. In this work, we propose a new learning framework, model-level dual learning, which takes duality of tasks into consideration while designing the architectures for the primal/dual models, and ties the model parameters that playing similar roles in the two tasks. We study both symmetric and asymmetric model-level dual learning. Our algorithms achieve significant improvements on neural machine translation and sentiment analysis.
****
Dynamical Isometry and a Mean Field Theory of CNNs: How to Train 10,000-Layer Vanilla Convolutional Neural Networks
Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, Jeffrey Pennington
In recent years, state-of-the-art methods in computer vision have utilized increasingly deep convolutional neural network architectures (CNNs), with some of the most successful models employing hundreds or even thousands of layers. A variety of pathologies such as vanishing/exploding gradients make training such deep networks challenging. While residual connections and batch normalization do enable training at these depths, it has remained unclear whether such specialized architecture designs are truly necessary to train deep CNNs. In this work, we demon

strate that it is possible to train vanilla CNNs with ten thousand layers or mor e simply by using an appropriate initialization scheme. We derive this initializ ation scheme theoretically by developing a mean field theory for signal propagat ion and by characterizing the conditions for dynamical isometry, the equilibrati on of singular values of the input-output Jacobian matrix. These conditions requ ire that the convolution operator be an orthogonal transformation in the sense t hat it is norm-preserving. We present an algorithm for generating such random in itial orthogonal convolution kernels and demonstrate empirically that they enabl e efficient training of extremely deep architectures.
****

Orthogonality-Promoting Distance Metric Learning: Convex Relaxation and Theoreti cal Analysis
Pengtao Xie, Wei Wu, Yichen Zhu, Eric Xing
Distance metric learning (DML), which learns a distance metric from labeled "sim ilar" and "dissimilar" data pairs, is widely utilized. Recently, several works i nvestigate orthogonality-promoting regularization (OPR), which encourages the pr ojection vectors in DML to be close to being orthogonal, to achieve three effect s: (1) high balancedness – achieving comparable performance on both frequent and infrequent classes; (2) high compactness – using a small number of projection v ectors to achieve a "good" metric; (3) good generalizability – alleviating overf itting to training data. While showing promising results, these approaches suffe r three problems. First, they involve solving non-convex optimization problems w here achieving the global optimal is NP-hard. Second, it lacks a theoretical und erstanding why OPR can lead to balancedness. Third, the current generalization e rror analysis of OPR is not directly on the regularizer. In this paper, we addre ss these three issues by (1) seeking convex relaxations of the original nonconve x problems so that the global optimal is guaranteed to be achievable; (2) provid ing a formal analysis on OPR’s capability of promoting balancedness; (3) providi ng a theoretical analysis that directly reveals the relationship between OPR and generalization performance. Experiments on various datasets demonstrate that ou r convex methods are more effective in promoting balancedness, compactness, and generalization, and are computationally more efficient, compared with the noncon vex methods.
****

Nonoverlap-Promoting Variable Selection
Pengtao Xie, Hongbao Zhang, Yichen Zhu, Eric Xing
Variable selection is a classic problem in machine learning (ML), widely used to find important explanatory factors, and improve generalization performance and interpretability of ML models. In this paper, we consider variable selection for models where multiple responses are to be predicted based on the same set of co variates. Since each response is relevant to a unique subset of covariates, we d esire the selected variables for different responses have small overlap. We prop ose a regularizer that simultaneously encourage orthogonality and sparsity, whic h jointly brings in an effect of reducing overlap. We apply this regularizer to four model instances and develop efficient algorithms to solve the regularized p roblems. We provide a formal analysis on why the proposed regularizer can reduce generalization error. Experiments on both simulation studies and real-world dat asets demonstrate the effectiveness of the proposed regularizer in selecting les s-overlapped variables and improving generalization performance.
****

Learning Semantic Representations for Unsupervised Domain Adaptation
Shaoan Xie, Zibin Zheng, Liang Chen, Chuan Chen
It is important to transfer the knowledge from label-rich source domain to unlab eled target domain due to the expensive cost of manual labeling efforts. Prior d omain adaptation methods address this problem through aligning the global distri bution statistics between source domain and target domain, but a drawback of pri or methods is that they ignore the semantic information contained in samples, e. g., features of backpacks in target domain might be mapped near features of cars in source domain. In this paper, we present moving semantic transfer network, w hich learn semantic representations for unlabeled target samples by aligning lab

eled source centroid and pseudo-labeled target centroid. Features in same class but different domains are expected to be mapped nearby, resulting in an improved target classification accuracy. Moving average centroid alignment is cautiously designed to compensate the insufficient categorical information within each mini batch. Experiments testify that our model yields state of the art results on standard datasets.

****

Rates of Convergence of Spectral Methods for Graphon Estimation

Jiaming Xu

This paper studies the problem of estimating the graphon function – a generative mechanism for a class of random graphs that are useful approximations to real networks. Specifically, a graph of n vertices is generated such that each pair of two vertices i and j are connected independently with probability $\rho_n \times f(x_i,x_j)$, where x_i is the unknown d-dimensional label of vertex i, f is an unknown symmetric function, and ρ_n, assumed to be $\Omega(\log n/n)$, is a scaling parameter characterizing the graph sparsity. The task is to estimate graphon f given the graph. Recent studies have identified the minimax optimal estimation error rate for $d=1$. However, there exists a wide gap between the known error rates of polynomial-time estimators and the minimax optimal error rate. We improve on the previously known error rates of polynomial-time estimators, by analyzing a spectral method, namely universal singular value thresholding (USVT) algorithm. When f belongs to either Hölder or Sobolev space with smoothness index α, we show the error rates of USVT are at most $(n\rho)^{ -2 \alpha / (2\alpha+d)}$. These error rates approach the minimax optimal error rate $\log (n\rho)/(n\rho)$ proved in prior work for $d=1$, as α increases, i.e., f becomes smoother. Furthermore, when f is analytic with infinitely many times differentiability, we show the error rate of USVT is at most $\log^d (n\rho)/(n\rho)$. When f is a step function which corresponds to the stochastic block model with k blocks for some k, the error rate of USVT is at most $k/(n\rho)$, which is larger than the minimax optimal error rate by at most a multiplicative factor $k/\log k$. This coincides with the computational gap observed in community detection. A key ingredient of our analysis is to derive the eigenvalue decaying rate of the edge probability matrix using piecewise polynomial approximations of the graphon function f.

****

Learning Registered Point Processes from Idiosyncratic Observations

Hongteng Xu, Lawrence Carin, Hongyuan Zha

A parametric point process model is developed, with modeling based on the assumption that sequential observations often share latent phenomena, while also possessing idiosyncratic effects. An alternating optimization method is proposed to learn a “registered” point process that accounts for shared structure, as well as “warping” functions that characterize idiosyncratic aspects of each observed sequence. Under reasonable constraints, in each iteration we update the sample-specific warping functions by solving a set of constrained nonlinear programming problems in parallel, and update the model by maximum likelihood estimation. The justifiability, complexity and robustness of the proposed method are investigated in detail, and the influence of sequence stitching on the learning results is examined empirically. Experiments on both synthetic and real-world data demonstrate that the method yields explainable point process models, achieving encouraging results compared to state-of-the-art methods.

****

Representation Learning on Graphs with Jumping Knowledge Networks

Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, Stefanie Jegelka

Recent deep learning approaches for representation learning on graphs follow a neighborhood aggregation procedure. We analyze some important properties of these models, and propose a strategy to overcome those. In particular, the range of "neighboring" nodes that a node’s representation draws from strongly depends on the graph structure, analogous to the spread of a random walk. To adapt to local neighborhood properties and tasks, we explore an architecture – jumping knowledg

e (JK) networks – that flexibly leverages, for each node, different neighborhood ranges to enable better structure-aware representation. In a number of experiments on social, bioinformatics and citation networks, we demonstrate that our model achieves state-of-the-art performance. Furthermore, combining the JK framework with models like Graph Convolutional Networks, GraphSAGE and Graph Attention Networks consistently improves those models' performance.

****

Learning to Explore via Meta-Policy Gradient

Tianbing Xu, Qiang Liu, Liang Zhao, Jian Peng

The performance of off-policy learning, including deep Q-learning and deep deterministic policy gradient (DDPG), critically depends on the choice of the exploration policy. Existing exploration methods are mostly based on adding noise to the on-going actor policy and can only explore local regions close to what the actor policy dictates. In this work, we develop a simple meta-policy gradient algorithm that allows us to adaptively learn the exploration policy in DDPG. Our algorithm allows us to train flexible exploration behaviors that are independent of the actor policy, yielding a global exploration that significantly speeds up the learning process. With an extensive study, we show that our method significantly improves the sample-efficiency of DDPG on a variety of reinforcement learning continuous control tasks.

****

Nonparametric Regression with Comparisons: Escaping the Curse of Dimensionality with Ordinal Information

Yichong Xu, Hariank Muthakana, Sivaraman Balakrishnan, Aarti Singh, Artur Dubrawski

In supervised learning, we leverage a labeled dataset to design methods for function estimation. In many practical situations, we are able to obtain alternative feedback, possibly at a low cost. A broad goal is to understand the usefulness of, and to design algorithms to exploit, this alternative feedback. We focus on a semi-supervised setting where we obtain additional ordinal (or comparison) information for potentially unlabeled samples. We consider ordinal feedback of varying qualities where we have either a perfect ordering of the samples, a noisy ordering of the samples or noisy pairwise comparisons between the samples. We provide a precise quantification of the usefulness of these types of ordinal feedback in non-parametric regression, showing that in many cases it is possible to accurately estimate an underlying function with a very small labeled set, effectively escaping the curse of dimensionality. We develop an algorithm called Ranking-Regression (RR) and analyze its accuracy as a function of size of the labeled and unlabeled datasets and various noise parameters. We also present lower bounds, that establish fundamental limits for the task and show that RR is optimal in a variety of settings. Finally, we present experiments that show the efficacy of RR and investigate its robustness to various sources of noise and model-misspecification.

****

Optimal Tuning for Divide-and-conquer Kernel Ridge Regression with Massive Data

Ganggang Xu, Zuofeng Shang, Guang Cheng

Divide-and-conquer is a powerful approach for large and massive data analysis. In the nonparameteric regression setting, although various theoretical frameworks have been established to achieve optimality in estimation or hypothesis testing, how to choose the tuning parameter in a practically effective way is still an open problem. In this paper, we propose a data-driven procedure based on divide-and-conquer for selecting the tuning parameters in kernel ridge regression by modifying the popular Generalized Cross-validation (GCV, Wahba, 1990). While the proposed criterion is computationally scalable for massive data sets, it is also shown under mild conditions to be asymptotically optimal in the sense that minimizing the proposed distributed-GCV (dGCV) criterion is equivalent to minimizing the true global conditional empirical loss of the averaged function estimator, extending the existing optimality results of GCV to the divide-and-conquer framework.

****

Continuous and Discrete-time Accelerated Stochastic Mirror Descent for Strongly Convex Functions
Pan Xu, Tianhao Wang, Quanquan Gu
We provide a second-order stochastic differential equation (SDE), which characterizes the continuous-time dynamics of accelerated stochastic mirror descent (ASMD) for strongly convex functions. This SDE plays a central role in designing new discrete-time ASMD algorithms via numerical discretization, and providing neat analyses of their convergence rates based on Lyapunov functions. Our results suggest that the only existing ASMD algorithm, namely, AC-SA proposed in Ghadimi & Lan (2012) is one instance of its kind, and we can actually derive new instances of ASMD with fewer tuning parameters. This sheds light on revisiting accelerated stochastic optimization through the lens of SDEs, which can lead to a better understanding of acceleration in stochastic optimization, as well as new simpler algorithms. Numerical experiments on both synthetic and real data support our theory.
****
A Semantic Loss Function for Deep Learning with Symbolic Knowledge
Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, Guy Broeck
This paper develops a novel methodology for using symbolic knowledge in deep learning. From first principles, we derive a semantic loss function that bridges between neural output vectors and logical constraints. This loss function captures how close the neural network is to satisfying the constraints on its output. An experimental evaluation shows that it effectively guides the learner to achieve (near-)state-of-the-art results on semi-supervised multi-class classification. Moreover, it significantly increases the ability of the neural network to predict structured objects, such as rankings and paths. These discrete concepts are tremendously difficult to learn, and benefit from a tight integration of deep learning and symbolic reasoning methods.
****
Causal Bandits with Propagating Inference
Akihiro Yabe, Daisuke Hatano, Hanna Sumita, Shinji Ito, Naonori Kakimura, Takuro Fukunaga, Ken-ichi Kawarabayashi
Bandit is a framework for designing sequential experiments, where a learner selects an arm $A \in \mathcal{A}$ and obtains an observation corresponding to A in each experiment. Theoretically, the tight regret lower-bound for the general bandit is polynomial with respect to the number of arms $|\mathcal{A}|$, and thus, to overcome this bound, the bandit problem with side-information is often considered. Recently, a bandit framework over a causal graph was introduced, where the structure of the causal graph is available as side-information and the arms are identified with interventions on the causal graph. Existing algorithms for causal bandit overcame the $\Omega(\sqrt{|\mathcal{A}|/T})$ simple-regret lower-bound; however, their algorithms work only when the interventions $\mathcal{A}$ are localized around a single node (i.e., an intervention propagates only to its neighbors). We then propose a novel causal bandit algorithm for an arbitrary set of interventions, which can propagate throughout the causal graph. We also show that it achieves $O(\sqrt{ \gamma^*\log(|\mathcal{A}|T) / T})$ regret bound, where γ^* is determined by using a causal graph structure. In particular, if the maximum in-degree of the causal graph is a constant, then $\gamma^* = O(N^2)$, where N is the number of nodes.
****
Active Learning with Logged Data
Songbai Yan, Kamalika Chaudhuri, Tara Javidi
We consider active learning with logged data, where labeled examples are drawn conditioned on a predetermined logging policy, and the goal is to learn a classifier on the entire population, not just conditioned on the logging policy. Prior work addresses this problem either when only logged data is available, or purely in a controlled random experimentation setting where the logged data is ignored. In this work, we combine both approaches to provide an algorithm that uses logged data to bootstrap and inform experimentation, thus achieving the best of both worlds. Our work is inspired by a connection between controlled random experim

entation and active learning, and modifies existing disagreement-based active le
arning algorithms to exploit logged data.
****

Binary Classification with Karmic, Threshold-Quasi-Concave Metrics
Bowei Yan, Sanmi Koyejo, Kai Zhong, Pradeep Ravikumar
Complex performance measures, beyond the popular measure of accuracy, are increa
singly being used in the context of binary classification. These complex perform
ance measures are typically not even decomposable, that is, the loss evaluated o
n a batch of samples cannot typically be expressed as a sum or average of losses
 evaluated at individual samples, which in turn requires new theoretical and met
hodological developments beyond standard treatments of supervised learning. In t
his paper, we advance this understanding of binary classification for complex pe
rformance measures by identifying two key properties: a so-called Karmic propert
y, and a more technical threshold-quasi-concavity property, which we show is mil
der than existing structural assumptions imposed on performance measures. Under
these properties, we show that the Bayes optimal classifier is a threshold funct
ion of the conditional probability of positive class. We then leverage this resu
lt to come up with a computationally practical plug-in classifier, via a novel t
hreshold estimator, and further, provide a novel statistical analysis of classif
ication error with respect to complex performance measures.
****

Characterizing and Learning Equivalence Classes of Causal DAGs under Interventio
ns
Karren Yang, Abigail Katcoff, Caroline Uhler
We consider the problem of learning causal DAGs in the setting where both observ
ational and interventional data is available. This setting is common in biology,
 where gene regulatory networks can be intervened on using chemical reagents or
gene deletions. Hauser & Buhlmann (2012) previously characterized the identifiab
ility of causal DAGs under perfect interventions, which eliminate dependencies b
etween targeted variables and their direct causes. In this paper, we extend thes
e identifiability results to general interventions, which may modify the depende
ncies between targeted variables and their causes without eliminating them. We d
efine and characterize the interventional Markov equivalence class that can be i
dentified from general (not necessarily perfect) intervention experiments. We al
so propose the first provably consistent algorithm for learning DAGs in this set
ting and evaluate our algorithm on simulated and biological datasets.
****

Dependent Relational Gamma Process Models for Longitudinal Networks
Sikun Yang, Heinz Koeppl
A probabilistic framework based on the covariate-dependent relational gamma proc
ess is developed to analyze relational data arising from longitudinal networks.
The proposed framework characterizes networked nodes by nonnegative node-group m
emberships, which allow each node to belong to multiple latent groups simultaneo
usly, and encodes edge probabilities between each pair of nodes using a Bernoull
i Poisson link to the embedded latent space. Within the latent space, our framew
ork models the birth and death dynamics of individual groups via a thinning func
tion. Our framework also captures the evolution of individual node-group members
hips over time using gamma Markov processes. Exploiting the recent advances in d
ata augmentation and marginalization techniques, a simple and efficient Gibbs sa
mpler is proposed for posterior computation. Experimental results on a simulatio
n study and three real-world temporal network data sets demonstrate the model’s
capability, competitive performance and scalability compared to state-of-the-art
 methods.
****

Goodness-of-Fit Testing for Discrete Distributions via Stein Discrepancy
Jiasen Yang, Qiang Liu, Vinayak Rao, Jennifer Neville
Recent work has combined Stein’s method with reproducing kernel Hilbert space th
eory to develop nonparametric goodness-of-fit tests for un-normalized probabilit
y distributions. However, the currently available tests apply exclusively to dis
tributions with smooth density functions. In this work, we introduce a kernelize

d Stein discrepancy measure for discrete spaces, and develop a nonparametric goo
dness-of-fit test for discrete distributions with intractable normalization cons
tants. Furthermore, we propose a general characterization of Stein operators tha
t encompasses both discrete and continuous distributions, providing a recipe for
 constructing new Stein operators. We apply the proposed goodness-of-fit test to
 three statistical models involving discrete distributions, and our experiments
show that the proposed test typically outperforms a two-sample test based on the
 maximum mean discrepancy.
****

Mean Field Multi-Agent Reinforcement Learning
Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, Jun Wang
Existing multi-agent reinforcement learning methods are limited typically to a s
mall number of agents. When the agent number increases largely, the learning bec
omes intractable due to the curse of the dimensionality and the exponential grow
th of agent interactions. In this paper, we present Mean Field Reinforcement Lea
rning where the interactions within the population of agents are approximated by
 those between a single agent and the average effect from the overall population
 or neighboring agents; the interplay between the two entities is mutually reinf
orced: the learning of the individual agent’s optimal policy depends on the dyna
mics of the population, while the dynamics of the population change according to
 the collective patterns of the individual policies. We develop practical mean f
ield Q-learning and mean field Actor-Critic algorithms and analyze the convergen
ce of the solution to Nash equilibrium. Experiments on Gaussian squeeze, Ising m
odel, and battle games justify the learning effectiveness of our mean field appr
oaches. In addition, we report the first result to solve the Ising model via mod
el-free reinforcement learning methods.
****

Yes, but Did It Work?: Evaluating Variational Inference
Yuling Yao, Aki Vehtari, Daniel Simpson, Andrew Gelman
While it’s always possible to compute a variational approximation to a posterior
 distribution, it can be difficult to discover problems with this approximation.
 We propose two diagnostic algorithms to alleviate this problem. The Pareto-smoo
thed importance sampling (PSIS) diagnostic gives a goodness of fit measurement f
or joint distributions, while simultaneously improving the error in the estimate
. The variational simulation-based calibration (VSBC) assesses the average perfo
rmance of point estimates.
****

Hierarchical Text Generation and Planning for Strategic Dialogue
Denis Yarats, Mike Lewis
End-to-end models for goal-orientated dialogue are challenging to train, because
 linguistic and strategic aspects are entangled in latent state vectors. We intr
oduce an approach to learning representations of messages in dialogues by maximi
zing the likelihood of subsequent sentences and actions, which decouples the sem
antics of the dialogue utterance from its linguistic realization. We then use th
ese latent sentence representations for hierarchical language generation, planni
ng and reinforcement learning. Experiments show that our approach increases the
end-task reward achieved by the model, improves the effectiveness of long-term p
lanning using rollouts, and allows self-play reinforcement learning to improve d
ecision making without diverging from human language. Our hierarchical latent-va
riable model outperforms previous work both linguistically and strategically.
****

Massively Parallel Algorithms and Hardness for Single-Linkage Clustering under $
\ell_p$ Distances
Grigory Yaroslavtsev, Adithya Vadapalli
We present first massively parallel (MPC) algorithms and hardness of approximati
on results for computing Single-Linkage Clustering of n input d-dimensional vect
ors under Hamming, ℓ_1, ℓ_2 and ℓ_∞ distances. All our algorit
hms run in O(log n) rounds of MPC for any fixed d and achieve (1+\epsilon)-appro
ximation for all distances (except Hamming for which we show an exact algorithm)
. We also show constant-factor inapproximability results for o(\log n)-round alg

orithms under standard MPC hardness assumptions (for sufficiently large dimensio n depending on the distance used). Efficiency of implementation of our algorithm s in Apache Spark is demonstrated through experiments on the largest available v ector datasets from the UCI machine learning repository exhibiting speedups of s everal orders of magnitude.
****

Communication-Computation Efficient Gradient Coding
Min Ye, Emmanuel Abbe
This paper develops coding techniques to reduce the running time of distributed learning tasks. It characterizes the fundamental tradeoff to compute gradients i n terms of three parameters: computation load, straggler tolerance and communica tion cost. It further gives an explicit coding scheme that achieves the optimal tradeoff based on recursive polynomial constructions, coding both across data su bsets and vector components. As a result, the proposed scheme allows to minimize the running time for gradient computations. Implementations are made on Amazon EC2 clusters using Python with mpi4py package. Results show that the proposed sc heme maintains the same generalization error while reducing the running time by $32%$ compared to uncoded schemes and $23%$ compared to prior coded schemes focu sing only on stragglers (Tandon et al., ICML 2017).
****

Variable Selection via Penalized Neural Network: a Drop-Out-One Loss Approach
Mao Ye, Yan Sun
We propose a variable selection method for high dimensional regression models, w hich allows for complex, nonlinear, and high-order interactions among variables. The proposed method approximates this complex system using a penalized neural n etwork and selects explanatory variables by measuring their utility in explainin g the variance of the response variable. This measurement is based on a novel st atistic called Drop-Out-One Loss. The proposed method also allows (overlapping) group variable selection. We prove that the proposed method can select relevant variables and exclude irrelevant variables with probability one as the sample si ze goes to infinity, which is referred to as the Oracle Property. Experimental r esults on simulated and real world datasets show the efficiency of our method in terms of variable selection and prediction accuracy.
****

Loss Decomposition for Fast Learning in Large Output Spaces
Ian En-Hsu Yen, Satyen Kale, Felix Yu, Daniel Holtmann-Rice, Sanjiv Kumar, Prade ep Ravikumar
For problems with large output spaces, evaluation of the loss function and its g radient are expensive, typically taking linear time in the size of the output sp ace. Recently, methods have been developed to speed up learning via efficient da ta structures for Nearest-Neighbor Search (NNS) or Maximum Inner-Product Search (MIPS). However, the performance of such data structures typically degrades in h igh dimensions. In this work, we propose a novel technique to reduce the intract able high dimensional search problem to several much more tractable lower dimens ional ones via dual decomposition of the loss function. At the same time, we dem onstrate guaranteed convergence to the original loss via a greedy message passin g procedure. In our experiments on multiclass and multilabel classification with hundreds of thousands of classes, as well as training skip-gram word embeddings with a vocabulary size of half a million, our technique consistently improves t he accuracy of search-based gradient approximation methods and outperforms sampl ing-based gradient approximation methods by a large margin.
****

Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates
Dong Yin, Yudong Chen, Ramchandran Kannan, Peter Bartlett
In this paper, we develop distributed optimization algorithms that are provably robust against Byzantine failures—arbitrary and potentially adversarial behavior , in distributed computing systems, with a focus on achieving optimal statistica l performance. A main result of this work is a sharp analysis of two robust dist ributed gradient descent algorithms based on median and trimmed mean operations, respectively. We prove statistical error rates for all of strongly convex, non-

strongly convex, and smooth non-convex population loss functions. In particular, these algorithms are shown to achieve order-optimal statistical error rates for strongly convex losses. To achieve better communication efficiency, we further propose a median-based distributed algorithm that is provably robust, and uses only one communication round. For strongly convex quadratic loss, we show that this algorithm achieves the same optimal error rate as the robust distributed gradient descent algorithms.
****
Semi-Implicit Variational Inference
Mingzhang Yin, Mingyuan Zhou
Semi-implicit variational inference (SIVI) is introduced to expand the commonly used analytic variational distribution family, by mixing the variational parameter with a flexible distribution. This mixing distribution can assume any density function, explicit or not, as long as independent random samples can be generated via reparameterization. Not only does SIVI expand the variational family to incorporate highly flexible variational distributions, including implicit ones that have no analytic density functions, but also sandwiches the evidence lower bound (ELBO) between a lower bound and an upper bound, and further derives an asymptotically exact surrogate ELBO that is amenable to optimization via stochastic gradient ascent. With a substantially expanded variational family and a novel optimization algorithm, SIVI is shown to closely match the accuracy of MCMC in inferring the posterior in a variety of Bayesian inference tasks.
****
Disentangled Sequential Autoencoder
Li Yingzhen, Stephan Mandt
We present a VAE architecture for encoding and generating high dimensional sequential data, such as video or audio. Our deep generative model learns a latent representation of the data which is split into a static and dynamic part, allowing us to approximately disentangle latent time-dependent features (dynamics) from features which are preserved over time (content). This architecture gives us partial control over generating content and dynamics by conditioning on either one of these sets of features. In our experiments on artificially generated cartoon video clips and voice recordings, we show that we can convert the content of a given sequence into another one by such content swapping. For audio, this allows us to convert a male speaker into a female speaker and vice versa, while for video we can separately manipulate shapes and dynamics. Furthermore, we give empirical evidence for the hypothesis that stochastic RNNs as latent state models are more efficient at compressing and generating long sequences than deterministic ones, which may be relevant for applications in video compression.
****
Probably Approximately Metric-Fair Learning
Gal Yona, Guy Rothblum
The seminal work of Dwork et al. [ITCS 2012] introduced a metric-based notion of individual fairness: given a task-specific similarity metric, their notion required that every pair of similar individuals should be treated similarly. In the context of machine learning, however, individual fairness does not generalize from a training set to the underlying population. We show that this can lead to computational intractability even for simple fair-learning tasks. With this motivation in mind, we introduce and study a relaxed notion of approximate metric-fairness: for a random pair of individuals sampled from the population, with all but a small probability of error, if they are similar then they should be treated similarly. We formalize the goal of achieving approximate metric-fairness simultaneously with best-possible accuracy as Probably Approximately Correct and Fair (PACF) Learning. We show that approximate metric-fairness does generalize, and leverage these generalization guarantees to construct polynomial-time PACF learning algorithms for the classes of linear and logistic predictors.
****
GAIN: Missing Data Imputation using Generative Adversarial Nets
Jinsung Yoon, James Jordon, Mihaela Schaar
We propose a novel method for imputing missing data by adapting the well-known G

enerative Adversarial Nets (GAN) framework. Accordingly, we call our method Gene rative Adversarial Imputation Nets (GAIN). The generator (G) observes some compo nents of a real data vector, imputes the missing components conditioned on what is actually observed, and outputs a completed vector. The discriminator (D) then takes a completed vector and attempts to determine which components were actual ly observed and which were imputed. To ensure that D forces G to learn the desir ed distribution, we provide D with some additional information in the form of a hint vector. The hint reveals to D partial information about the missingness of the original sample, which is used by D to focus its attention on the imputation quality of particular components. This hint ensures that G does in fact learn t o generate according to the true data distribution. We tested our method on vari ous datasets and found that GAIN significantly outperforms state-of-the-art impu tation methods.

****

RadialGAN: Leveraging multiple datasets to improve target-specific predictive mo dels using Generative Adversarial Networks

Jinsung Yoon, James Jordon, Mihaela Schaar

Training complex machine learning models for prediction often requires a large a mount of data that is not always readily available. Leveraging these external da tasets from related but different sources is therefore an important task if good predictive models are to be built for deployment in settings where data can be rare. In this paper we propose a novel approach to the problem in which we use m ultiple GAN architectures to learn to translate from one dataset to another, the reby allowing us to effectively enlarge the target dataset, and therefore learn better predictive models than if we simply used the target dataset. We show the utility of such an approach, demonstrating that our method improves the predicti on performance on the target domain over using just the target dataset and also show that our framework outperforms several other benchmarks on a collection of real-world medical datasets.

****

GraphRNN: Generating Realistic Graphs with Deep Auto-regressive Models

Jiaxuan You, Rex Ying, Xiang Ren, William Hamilton, Jure Leskovec

Modeling and generating graphs is fundamental for studying networks in biology, engineering, and social sciences. However, modeling complex distributions over g raphs and then efficiently sampling from these distributions is challenging due to the non-unique, high-dimensional nature of graphs and the complex, non-local dependencies that exist between edges in a given graph. Here we propose GraphRNN , a deep autoregressive model that addresses the above challenges and approximat es any distribution of graphs with minimal assumptions about their structure. Gr aphRNN learns to generate graphs by training on a representative set of graphs a nd decomposes the graph generation process into a sequence of node and edge form ations, conditioned on the graph structure generated so far. In order to quantit atively evaluate the performance of GraphRNN, we introduce a benchmark suite of datasets, baselines and novel evaluation metrics based on Maximum Mean Discrepan cy, which measure distances between sets of graphs. Our experiments show that Gr aphRNN significantly outperforms all baselines, learning to generate diverse gra phs that match the structural characteristics of a target set, while also scalin g to graphs 50 times larger than previous deep models.

****

An Efficient Semismooth Newton based Algorithm for Convex Clustering

Yancheng Yuan, Defeng Sun, Kim-Chuan Toh

Clustering is a fundamental problem in unsupervised learning. Popular methods li ke K-means, may suffer from instability as they are prone to get stuck in its lo cal minima. Recently, the sumof-norms (SON) model (also known as clustering path), which is a convex relaxation of hierarchical clustering model, has been propo sed in (Lindsten et al., 2011) and (Hocking et al., 2011). Although numerical al gorithms like alternating direction method of multipliers (ADMM) and alternating minimization algorithm (AMA) have been proposed to solve convex clustering mode l (Chi & Lange, 2015), it is known to be very challenging to solve large-scale p roblems. In this paper, we propose a semismooth Newton based augmented Lagrangia

n method for large-scale convex clustering problems. Extensive numerical experim ents on both simulated and real data demonstrate that our algorithm is highly ef ficient and robust for solving large-scale problems. Moreover, the numerical res ults also show the superior performance and scalability of our algorithm compari ng to existing first-order methods.
****

A Conditional Gradient Framework for Composite Convex Minimization with Applicat ions to Semidefinite Programming
Alp Yurtsever, Olivier Fercoq, Francesco Locatello, Volkan Cevher
We propose a conditional gradient framework for a composite convex minimization template with broad applications. Our approach combines smoothing and homotopy t echniques under the CGM framework, and provably achieves the optimal convergence rate. We demonstrate that the same rate holds if the linear subproblems are sol ved approximately with additive or multiplicative error. In contrast with the re levant work, we are able to characterize the convergence when the non-smooth ter m is an indicator function. Specific applications of our framework include the n on-smooth minimization, semidefinite programming, and minimization with linear i nclusion constraints over a compact domain. Numerical evidence demonstrates the benefits of our framework.
****

Policy Optimization as Wasserstein Gradient Flows
Ruiyi Zhang, Changyou Chen, Chunyuan Li, Lawrence Carin
Policy optimization is a core component of reinforcement learning (RL), and most existing RL methods directly optimize parameters of a policy based on maximizin g the expected total reward, or its surrogate. Though often achieving encouragin g empirical success, its correspondence to policy-distribution optimization has been unclear mathematically. We place policy optimization into the space of prob ability measures, and interpret it as Wasserstein gradient flows. On the probabi lity-measure space, under specified circumstances, policy optimization becomes c onvex in terms of distribution optimization. To make optimization feasible, we d evelop efficient algorithms by numerically solving the corresponding discrete gr adient flows. Our technique is applicable to several RL settings, and is related to many state-of-the-art policy-optimization algorithms. Specifically, we defin e gradient flows on both the parameter-distribution space and policy-distributio n space, leading to what we term indirect-policy and direct-policy learning fram eworks, respectively. Extensive experiments verify the effectiveness of our fram ework, often obtaining better performance compared to related algorithms.
****

Problem Dependent Reinforcement Learning Bounds Which Can Identify Bandit Struct ure in MDPs
Andrea Zanette, Emma Brunskill
In order to make good decision under uncertainty an agent must learn from observ ations. To do so, two of the most common frameworks are Contextual Bandits and M arkov Decision Processes (MDPs). In this paper, we study whether there exist alg orithms for the more general framework (MDP) which automatically provide the bes t performance bounds for the specific problem at hand without user intervention and without modifying the algorithm. In particular, it is found that a very mino r variant of a recently proposed reinforcement learning algorithm for MDPs alrea dy matches the best possible regret bound $\tilde O (\sqrt{SAT})$ in the dominan t term if deployed on a tabular Contextual Bandit problem despite the agent bein g agnostic to such setting.
****

Fast and Sample Efficient Inductive Matrix Completion via Multi-Phase Procrustes Flow
Xiao Zhang, Simon Du, Quanquan Gu
We revisit the inductive matrix completion problem that aims to recover a rank-$ r$ matrix with ambient dimension d given n features as the side prior inform ation. The goal is to make use of the known n features to reduce sample and co mputational complexities. We present and analyze a new gradient-based non-convex optimization algorithm that converges to the true underlying matrix at a linear

rate with sample complexity only linearly depending on n and logarithmically depending on d. To the best of our knowledge, all previous algorithms either have a quadratic dependency on the number of features in sample complexity or a sub-linear computational convergence rate. In addition, we provide experiments on both synthetic and real world data to demonstrate the effectiveness of our proposed algorithm.

****

Large-Scale Sparse Inverse Covariance Estimation via Thresholding and Max-Det Matrix Completion

Richard Zhang, Salar Fattahi, Somayeh Sojoudi

The sparse inverse covariance estimation problem is commonly solved using an ℓ_{1}-regularized Gaussian maximum likelihood estimator known as “graphical lasso”, but its computational cost becomes prohibitive for large data sets. A recently line of results showed{–}under mild assumptions{–}that the graphical lasso estimator can be retrieved by soft-thresholding the sample covariance matrix and solving a maximum determinant matrix completion (MDMC) problem. This paper proves an extension of this result, and describes a Newton-CG algorithm to efficiently solve the MDMC problem. Assuming that the thresholded sample covariance matrix is sparse with a sparse Cholesky factorization, we prove that the algorithm converges to an ϵ-accurate solution in $O(n\log(1/\epsilon))$ time and $O(n)$ memory. The algorithm is highly efficient in practice: we solve the associated MDMC problems with as many as 200,000 variables to 7-9 digits of accuracy in less than an hour on a standard laptop computer running MATLAB.

****

High Performance Zero-Memory Overhead Direct Convolutions

Jiyuan Zhang, Franz Franchetti, Tze Meng Low

The computation of convolution layers in deep neural networks typically rely on high performance routines that trade space for time by using additional memory (either for packing purposes or required as part of the algorithm) to improve performance. The problems with such an approach are two-fold. First, these routines incur additional memory overhead which reduces the overall size of the network that can fit on embedded devices with limited memory capacity. Second, these high performance routines were not optimized for performing convolution, which means that the performance obtained is usually less than conventionally expected. In this paper, we demonstrate that direct convolution, when implemented correctly, eliminates all memory overhead, and yields performance that is between 10% to 400% times better than existing high performance implementations of convolution layers on conventional and embedded CPU architectures. We also show that a high performance direct convolution exhibits better scaling performance, i.e. suffers less performance drop, when increasing the number of threads.

****

Safe Element Screening for Submodular Function Minimization

Weizhong Zhang, Bin Hong, Lin Ma, Wei Liu, Tong Zhang

Submodular functions are discrete analogs of convex functions, which have applications in various fields, including machine learning and computer vision. However, in large-scale applications, solving Submodular Function Minimization (SFM) problems remains challenging. In this paper, we make the first attempt to extend the emerging technique named screening in large-scale sparse learning to SFM for accelerating its optimization process. We first conduct a careful studying of the relationships between SFM and the corresponding convex proximal problems, as well as the accurate primal optimum estimation of the proximal problems. Relying on this study, we subsequently propose a novel safe screening method to quickly identify the elements guaranteed to be included (we refer to them as active) or excluded (inactive) in the final optimal solution of SFM during the optimization process. By removing the inactive elements and fixing the active ones, the problem size can be dramatically reduced, leading to great savings in the computational cost without sacrificing any accuracy. To the best of our knowledge, the proposed method is the first screening method in the fields of SFM and even combinatorial optimization, thus pointing out a new direction for accelerating SFM algorithms. Experiment results on both synthetic and real datasets demonstrate the

significant speedups gained by our approach.
****
Improving the Privacy and Accuracy of ADMM-Based Distributed Algorithms
Xueru Zhang, Mohammad Mahdi Khalili, Mingyan Liu
Alternating direction method of multiplier (ADMM) is a popular method used to design distributed versions of a machine learning algorithm, whereby local computations are performed on local data with the output exchanged among neighbors in an iterative fashion. During this iterative process the leakage of data privacy arises. A differentially private ADMM was proposed in prior work (Zhang & Zhu, 2017) where only the privacy loss of a single node during one iteration was bounded, a method that makes it difficult to balance the tradeoff between the utility attained through distributed computation and privacy guarantees when considering the total privacy loss of all nodes over the entire iterative process. We propose a perturbation method for ADMM where the perturbed term is correlated with the penalty parameters; this is shown to improve the utility and privacy simultaneously. The method is based on a modified ADMM where each node independently determines its own penalty parameter in every iteration and decouples it from the dual updating step size. The condition for convergence of the modified ADMM and the lower bound on the convergence rate are also derived.
****
Stabilizing Gradients for Deep Neural Networks via Efficient SVD Parameterization
Jiong Zhang, Qi Lei, Inderjit Dhillon
Vanishing and exploding gradients are two of the main obstacles in training deep neural networks, especially in capturing long range dependencies in recurrent neural networks (RNNs). In this paper, we present an efficient parametrization of the transition matrix of an RNN that allows us to stabilize the gradients that arise in its training. Specifically, we parameterize the transition matrix by its singular value decomposition (SVD), which allows us to explicitly track and control its singular values. We attain efficiency by using tools that are common in numerical linear algebra, namely Householder reflectors for representing the orthogonal matrices that arise in the SVD. By explicitly controlling the singular values, our proposed Spectral-RNN method allows us to easily solve the exploding gradient problem and we observe that it empirically solves the vanishing gradient issue to a large extent. We note that the SVD parameterization can be used for any rectangular weight matrix, hence it can be easily extended to any deep neural network, such as a multi-layer perceptron. Theoretically, we demonstrate that our parameterization does not lose any expressive power, and show how it potentially makes the optimization process easier. Our extensive experimental results also demonstrate that the proposed framework converges faster, and has good generalization, especially in capturing long range dependencies, as shown on the synthetic addition and copy tasks, as well as on MNIST and Penn Tree Bank data sets.
****
Learning Long Term Dependencies via Fourier Recurrent Units
Jiong Zhang, Yibo Lin, Zhao Song, Inderjit Dhillon
It is a known fact that training recurrent neural networks for tasks that have long term dependencies is challenging. One of the main reasons is the vanishing or exploding gradient problem, which prevents gradient information from propagating to early layers. In this paper we propose a simple recurrent architecture, the Fourier Recurrent Unit (FRU), that stabilizes the gradients that arise in its training while giving us stronger expressive power. Specifically, FRU summarizes the hidden states $h^{(t)}$ along the temporal dimension with Fourier basis functions. This allows gradients to easily reach any layer due to FRU’s residual learning structure and the global support of trigonometric functions. We show that FRU has gradient lower and upper bounds independent of temporal dimension. We also show the strong expressivity of sparse Fourier basis, from which FRU obtains its strong expressive power. Our experimental study also demonstrates that with fewer parameters the proposed architecture outperforms other recurrent architectures on many tasks.

****
Tropical Geometry of Deep Neural Networks
Liwen Zhang, Gregory Naitzat, Lek-Heng Lim
We establish, for the first time, explicit connections between feedforward neural networks with ReLU activation and tropical geometry — we show that the family of such neural networks is equivalent to the family of tropical rational maps. Among other things, we deduce that feedforward ReLU neural networks with one hidden layer can be characterized by zonotopes, which serve as building blocks for deeper networks; we relate decision boundaries of such neural networks to tropical hypersurfaces, a major object of study in tropical geometry; and we prove that linear regions of such neural networks correspond to vertices of polytopes associated with tropical rational functions. An insight from our tropical formulation is that a deeper network is exponentially more expressive than a shallow network.
****
Deep Bayesian Nonparametric Tracking
Aonan Zhang, John Paisley
Time-series data often exhibit irregular behavior, making them hard to analyze and explain with a simple dynamic model. For example, information in social networks may show change-point-like bursts that then diffuse with smooth dynamics. Powerful models such as deep neural networks learn smooth functions from data, but are not as well-suited (in off-the-shelf form) for discovering and explaining sparse, discrete and bursty dynamic patterns. Bayesian models can do this well by encoding the appropriate probabilistic assumptions in the model prior. We propose an integration of Bayesian nonparametric methods within deep neural networks for modeling irregular patterns in time-series data. We use a Bayesian nonparametrics to model change-point behavior in time, and a deep neural network to model nonlinear latent space dynamics. We compare with a non-deep linear version of the model also proposed here. Empirical evaluations demonstrates improved performance and interpretable results when tracking stock prices and Twitter trends.
****
Composable Planning with Attributes
Amy Zhang, Sainbayar Sukhbaatar, Adam Lerer, Arthur Szlam, Rob Fergus
The tasks that an agent will need to solve often are not known during training. However, if the agent knows which properties of the environment are important then, after learning how its actions affect those properties, it may be able to use this knowledge to solve complex tasks without training specifically for them. Towards this end, we consider a setup in which an environment is augmented with a set of user defined attributes that parameterize the features of interest. We propose a method that learns a policy for transitioning between “nearby” sets of attributes, and maintains a graph of possible transitions. Given a task at test time that can be expressed in terms of a target set of attributes, and a current state, our model infers the attributes of the current state and searches over paths through attribute space to get a high level plan, and then uses its low level policy to execute the plan. We show in 3D block stacking, grid-world games, and StarCraft that our model is able to generalize to longer, more complex tasks at test time by composing simpler learned policies.
****
Noisy Natural Gradient as Variational Inference
Guodong Zhang, Shengyang Sun, David Duvenaud, Roger Grosse
Variational Bayesian neural nets combine the flexibility of deep learning with Bayesian uncertainty estimation. Unfortunately, there is a tradeoff between cheap but simple variational families (e.g. fully factorized) or expensive and complicated inference procedures. We show that natural gradient ascent with adaptive weight noise implicitly fits a variational posterior to maximize the evidence lower bound (ELBO). This insight allows us to train full-covariance, fully factorized, or matrix-variate Gaussian variational posteriors using noisy versions of natural gradient, Adam, and K-FAC, respectively, making it possible to scale up to modern-size ConvNets. On standard regression benchmarks, our noisy K-FAC algorithm makes better predictions and matches Hamiltonian Monte Carlo’s predictive va

riances better than existing methods. Its improved uncertainty estimates lead to more efficient exploration in active learning, and intrinsic motivation for reinforcement learning.
****

A Primal-Dual Analysis of Global Optimality in Nonconvex Low-Rank Matrix Recovery
Xiao Zhang, Lingxiao Wang, Yaodong Yu, Quanquan Gu
We propose a primal-dual based framework for analyzing the global optimality of nonconvex low-rank matrix recovery. Our analysis are based on the restricted strongly convex and smooth conditions, which can be verified for a broad family of loss functions. In addition, our analytic framework can directly handle the widely-used incoherence constraints through the lens of duality. We illustrate the applicability of the proposed framework to matrix completion and one-bit matrix completion, and prove that all these problems have no spurious local minima. Our results not only improve the sample complexity required for characterizing the global optimality of matrix completion, but also resolve an open problem in Ge et al. (2017) regarding one-bit matrix completion. Numerical experiments show that primal-dual based algorithm can successfully recover the global optimum for various low-rank problems.
****

Fully Decentralized Multi-Agent Reinforcement Learning with Networked Agents
Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, Tamer Basar
We consider the fully decentralized multi-agent reinforcement learning (MARL) problem, where the agents are connected via a time-varying and possibly sparse communication network. Specifically, we assume that the reward functions of the agents might correspond to different tasks, and are only known to the corresponding agent. Moreover, each agent makes individual decisions based on both the information observed locally and the messages received from its neighbors over the network. To maximize the globally averaged return over the network, we propose two fully decentralized actor-critic algorithms, which are applicable to large-scale MARL problems in an online fashion. Convergence guarantees are provided when the value functions are approximated within the class of linear functions. Our work appears to be the first theoretical study of fully decentralized MARL algorithms for networked agents that use function approximation.
****

Dynamic Regret of Strongly Adaptive Methods
Lijun Zhang, Tianbao Yang, jin, Zhi-Hua Zhou
To cope with changing environments, recent developments in online learning have introduced the concepts of adaptive regret and dynamic regret independently. In this paper, we illustrate an intrinsic connection between these two concepts by showing that the dynamic regret can be expressed in terms of the adaptive regret and the functional variation. This observation implies that strongly adaptive algorithms can be directly leveraged to minimize the dynamic regret. As a result, we present a series of strongly adaptive algorithms that have small dynamic regrets for convex functions, exponentially concave functions, and strongly convex functions, respectively. To the best of our knowledge, this is the first time that exponential concavity is utilized to upper bound the dynamic regret. Moreover, all of those adaptive algorithms do not need any prior knowledge of the functional variation, which is a significant advantage over previous specialized methods for minimizing dynamic regret.
****

Inter and Intra Topic Structure Learning with Word Embeddings
He Zhao, Lan Du, Wray Buntine, Mingyuan Zhou
One important task of topic modeling for text analysis is interpretability. By discovering structured topics one is able to yield improved interpretability as well as modeling accuracy. In this paper, we propose a novel topic model with a deep structure that explores both inter-topic and intra-topic structures informed by word embeddings. Specifically, our model discovers inter topic structures in the form of topic hierarchies and discovers intra topic structures in the form of sub-topics, each of which is informed by word embeddings and captures a fine-

grained thematic aspect of a normal topic. Extensive experiments demonstrate tha t our model achieves the state-of-the-art performance in terms of perplexity, do cument classification, and topic quality. Moreover, with topic hierarchies and s ub-topics, the topics discovered in our model are more interpretable, providing an illuminating means to understand text data.
****
Adversarially Regularized Autoencoders
Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander Rush, Yann LeCun
Deep latent variable models, trained using variational autoencoders or generativ e adversarial networks, are now a key technique for representation learning of c ontinuous structures. However, applying similar methods to discrete structures, such as text sequences or discretized images, has proven to be more challenging. In this work, we propose a more flexible method for training deep latent variab le models of discrete structures. Our approach is based on the recently proposed Wasserstein Autoencoder (WAE) which formalizes adversarial autoencoders as an o ptimal transport problem. We first extend this framework to model discrete seque nces, and then further explore different learned priors targeting a controllable representation. Unlike many other latent variable generative models for text, t his adversarially regularized autoencoder (ARAE) allows us to generate fluent te xtual outputs as well as perform manipulations in the latent space to induce cha nge in the output space. Finally we show that the latent representation can be t rained to perform unaligned textual style transfer, giving improvements both in automatic measures and human evaluation.
****
MSplit LBI: Realizing Feature Selection and Dense Estimation Simultaneously in F ew-shot and Zero-shot Learning
Bo Zhao, Xinwei Sun, Yanwei Fu, Yuan Yao, Yizhou Wang
It is one typical and general topic of learning a good embedding model to effici ently learn the representation coefficients between two spaces/subspaces. To sol ve this task, L_{1} regularization is widely used for the pursuit of feature s election and avoiding overfitting, and yet the sparse estimation of features in L_{1} regularization may cause the underfitting of training data. L_{2} regu larization is also frequently used, but it is a biased estimator. In this paper, we propose the idea that the features consist of three orthogonal parts, namely sparse strong signals, dense weak signals and random noise, in which both stron g and weak signals contribute to the fitting of data. To facilitate such novel d ecomposition, MSplit LBI is for the first time proposed to realize feature selec tion and dense estimation simultaneously. We provide theoretical and simulationa l verification that our method exceeds L_{1} and L_{2} regularization, and e xtensive experimental results show that our method achieves state-of-the-art per formance in the few-shot and zero-shot learning.
****
Composite Marginal Likelihood Methods for Random Utility Models
Zhibing Zhao, Lirong Xia
We propose a novel and flexible rank-breaking-then-composite-marginal-likelihood (RBCML) framework for learning random utility models (RUMs), which include the Plackett-Luce model. We characterize conditions for the objective function of RB CML to be strictly log-concave by proving that strict log-concavity is preserved under convolution and marginalization. We characterize necessary and sufficient conditions for RBCML to satisfy consistency and asymptotic normality. Experimen ts on synthetic data show that RBCML for Gaussian RUMs achieves better statistic al efficiency and computation efficiency than the state-of-the-art algorithm and our RBCML for the Plackett-Luce model provides flexible tradeoffs between runni ng time and statistical efficiency.
****
Lightweight Stochastic Optimization for Minimizing Finite Sums with Infinite Dat a
Shuai Zheng, James Tin-Yau Kwok
Variance reduction has been commonly used in stochastic optimization. It relies crucially on the assumption that the data set is finite. However, when the data

are imputed with random noise as in data augmentation, the perturbed data set be comes essentially infinite. Recently, the stochastic MISO (S-MISO) algorithm is introduced to address this expected risk minimization problem. Though it converg es faster than SGD, a significant amount of memory is required. In this paper, w e propose two SGD-like algorithms for expected risk minimization with random per turbation, namely, stochastic sample average gradient (SSAG) and stochastic SAGA (S-SAGA). The memory cost of SSAG does not depend on the sample size, while tha t of S-SAGA is the same as those of variance reduction methods on unperturbed da ta. Theoretical analysis and experimental results on logistic regression and AUC maximization show that SSAG has faster convergence rate than SGD with comparabl e space requirement while S-SAGA outperforms S-MISO in terms of both iteration c omplexity and storage.

****

A Robust Approach to Sequential Information Theoretic Planning

Sue Zheng, Jason Pacheco, John Fisher

In many sequential planning applications a natural approach to generating high q uality plans is to maximize an information reward such as mutual information (MI). Unfortunately, MI lacks a closed form in all but trivial models, and so must be estimated. In applications where the cost of plan execution is expensive, one desires planning estimates which admit theoretical guarantees. Through the use of robust M-estimators we obtain bounds on absolute deviation of estimated MI. M oreover, we propose a sequential algorithm which integrates inference and planni ng by maximally reusing particles in each stage. We validate the utility of usin g robust estimators in the sequential approach on a Gaussian Markov Random Field wherein information measures have a closed form. Lastly, we demonstrate the ben efits of our integrated approach in the context of sequential experiment design for inferring causal regulatory networks from gene expression levels. Our method shows improvements over a recent method which selects intervention experiments based on the same MI objective.

****

Revealing Common Statistical Behaviors in Heterogeneous Populations

Andrey Zhitnikov, Rotem Mulayoff, Tomer Michaeli

In many areas of neuroscience and biological data analysis, it is desired to rev eal common patterns among a group of subjects. Such analyses play important role s e.g., in detecting functional brain networks from fMRI scans and in identifyin g brain regions which show increased activity in response to certain stimuli. Gr oup level techniques usually assume that all subjects in the group behave accord ing to a single statistical model, or that deviations from the common model have simple parametric forms. Therefore, complex subject-specific deviations from th e common model severely impair the performance of such methods. In this paper, w e propose nonparametric algorithms for estimating the common covariance matrix a nd the common density function of several variables in a heterogeneous group of subjects. Our estimates converge to the true model as the number of subjects ten ds to infinity, under very mild conditions. We illustrate the effectiveness of o ur methods through extensive simulations as well as on real-data from fMRI scans and from arterial blood pressure and photoplethysmogram measurements.

****

Understanding Generalization and Optimization Performance of Deep CNNs

Pan Zhou, Jiashi Feng

This work aims to provide understandings on the remarkable success of deep convo lutional neural networks (CNNs) by theoretically analyzing their generalization performance and establishing optimization guarantees for gradient descent based training algorithms. Specifically, for a CNN model consisting of l convolution al layers and one fully connected layer, we prove that its generalization error is bounded by $\mathcal{O}(\sqrt{\theta\widetilde{\varrho}/n})$ where θ d enotes freedom degree of the network parameters and $\widetilde{\varrho}=\mathca l{O}(\log(\prod_{i=1}^{l}b_{i} (k_{i}-s_{i}+1)/p)+\log(b_{l+1}))$ encapsulates a rchitecture parameters including the kernel size k_{i}, stride s_{i}, poolin g size p and parameter magnitude b_{i}. To our best knowledge, this is the f irst generalization bound that only depends on $\mathcal{O}(\log(\prod_{i=1}^{l+

1}b_{i}))$, tighter than existing ones that all involve an exponential term like $\mathcal{O}(\prod_{i=1}^{l+1}b_{i})$. Besides, we prove that for an arbitrary gradient descent algorithm, the computed approximate stationary point by minimizing empirical risk is also an approximate stationary point to the population risk. This well explains why gradient descent training algorithms usually perform sufficiently well in practice. Furthermore, we prove the one-to-one correspondence and convergence guarantees for the non-degenerate stationary points between the empirical and population risks. It implies that the computed local minimum for the empirical risk is also close to a local minimum for the population risk, thus ensuring that the optimized CNN model well generalizes to new data.
****

Distributed Asynchronous Optimization with Unbounded Delays: How Slow Can You Go?
Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Peter Glynn, Yinyu Ye, Li-Jia Li, Li Fei-Fei
One of the most widely used optimization methods for large-scale machine learning problems is distributed asynchronous stochastic gradient descent (DASGD). However, a key issue that arises here is that of delayed gradients: when a “worker” node asynchronously contributes a gradient update to the “master”, the global model parameter may have changed, rendering this information stale. In massively parallel computing grids, these delays can quickly add up if the computational throughput of a node is saturated, so the convergence of DASGD is uncertain under these conditions. Nevertheless, by using a judiciously chosen quasilinear step-size sequence, we show that it is possible to amortize these delays and achieve global convergence with probability 1, even when the delays grow at a polynomial rate. In this way, our results help reaffirm the successful application of DASGD to large-scale optimization problems.
****

A Simple Stochastic Variance Reduced Algorithm with Fast Convergence Rates
Kaiwen Zhou, Fanhua Shang, James Cheng
Recent years have witnessed exciting progress in the study of stochastic variance reduced gradient methods (e.g., SVRG, SAGA), their accelerated variants (e.g, Katyusha) and their extensions in many different settings (e.g., online, sparse, asynchronous, distributed). Among them, accelerated methods enjoy improved convergence rates but have complex coupling structures, which makes them hard to be extended to more settings (e.g., sparse and asynchronous) due to the existence of perturbation. In this paper, we introduce a simple stochastic variance reduced algorithm (MiG), which enjoys the best-known convergence rates for both strongly convex and non-strongly convex problems. Moreover, we also present its efficient sparse and asynchronous variants, and theoretically analyze its convergence rates in these settings. Finally, extensive experiments for various machine learning problems such as logistic regression are given to illustrate the practical improvement in both serial and asynchronous settings.
****

Stochastic Variance-Reduced Cubic Regularized Newton Methods
Dongruo Zhou, Pan Xu, Quanquan Gu
We propose a stochastic variance-reduced cubic regularized Newton method (SVRC) for non-convex optimization. At the core of our algorithm is a novel semi-stochastic gradient along with a semi-stochastic Hessian, which are specifically designed for cubic regularization method. We show that our algorithm is guaranteed to converge to an $(\epsilon,\sqrt{\epsilon})$-approximate local minimum within $\tilde{O}(n^{4/5}/\epsilon^{3/2})$ second-order oracle calls, which outperforms the state-of-the-art cubic regularization algorithms including subsampled cubic regularization. Our work also sheds light on the application of variance reduction technique to high-order non-convex optimization methods. Thorough experiments on various non-convex optimization problems support our theory.
****

Racing Thompson: an Efficient Algorithm for Thompson Sampling with Non-conjugate Priors
Yichi Zhou, Jun Zhu, Jingwei Zhuo

Thompson sampling has impressive empirical performance for many multi-armed band it problems. But current algorithms for Thompson sampling only work for the case of conjugate priors since they require to perform online Bayesian posterior inf erence, which is a difficult task when the prior is not conjugate. In this paper , we propose a novel algorithm for Thompson sampling which only requires to draw samples from a tractable proposal distribution. So our algorithm is efficient e ven when the prior is non-conjugate. To do this, we reformulate Thompson samplin g as an optimization proplem via the Gumbel-Max trick. After that we construct a set of random variables and our goal is to identify the one with highest mean w hich is an instance of best arm identification problems. Finally, we solve it wi th techniques in best arm identification. Experiments show that our algorithm wo rks well in practice.

****

Distributed Nonparametric Regression under Communication Constraints

Yuancheng Zhu, John Lafferty

This paper studies the problem of nonparametric estimation of a smooth function with data distributed across multiple machines. We assume an independent sample from a white noise model is collected at each machine, and an estimator of the u nderlying true function needs to be constructed at a central machine. We place l imits on the number of bits that each machine can use to transmit information to the central machine. Our results give both asymptotic lower bounds and matching upper bounds on the statistical risk under various settings. We identify three regimes, depending on the relationship among the number of machines, the size of data available at each machine, and the communication budget. When the communic ation budget is small, the statistical risk depends solely on this communication bottleneck, regardless of the sample size. In the regime where the communicatio n budget is large, the classic minimax risk in the non-distributed estimation se tting is recovered. In an intermediate regime, the statistical risk depends on b oth the sample size and the communication budget.

****

Message Passing Stein Variational Gradient Descent

Jingwei Zhuo, Chang Liu, Jiaxin Shi, Jun Zhu, Ning Chen, Bo Zhang

Stein variational gradient descent (SVGD) is a recently proposed particle-based Bayesian inference method, which has attracted a lot of interest due to its rema rkable approximation ability and particle efficiency compared to traditional var iational inference and Markov Chain Monte Carlo methods. However, we observed th at particles of SVGD tend to collapse to modes of the target distribution, and t his particle degeneracy phenomenon becomes more severe with higher dimensions. O ur theoretical analysis finds out that there exists a negative correlation betwe en the dimensionality and the repulsive force of SVGD which should be blamed for this phenomenon. We propose Message Passing SVGD (MP-SVGD) to solve this proble m. By leveraging the conditional independence structure of probabilistic graphic al models (PGMs), MP-SVGD converts the original high-dimensional global inferenc e problem into a set of local ones over the Markov blanket with lower dimensions . Experimental results show its advantages of preventing vanishing repulsive for ce in high-dimensional space over SVGD, and its particle efficiency and approxim ation flexibility over other inference methods on graphical models.

****

Stochastic Variance-Reduced Hamilton Monte Carlo Methods

Difan Zou, Pan Xu, Quanquan Gu

We propose a fast stochastic Hamilton Monte Carlo (HMC) method, for sampling fro m a smooth and strongly log-concave distribution. At the core of our proposed me thod is a variance reduction technique inspired by the recent advance in stochas tic optimization. We show that, to achieve ϵ accuracy in 2-Wasserstein distance, our algorithm achieves $\tilde O\big(n+\kappa^{2}d^{1/2}/\epsilon+\kap pa^{4/3}d^{1/3}n^{2/3}/\epsilon^{2/3}\big)$ gradient complexity (i.e., number of component gradient evaluations), which outperforms the state-of-the-art HMC and stochastic gradient HMC methods in a wide regime. We also extend our algorithm for sampling from smooth and general log-concave distributions, and prove the co rresponding gradient complexity as well. Experiments on both synthetic and real

data demonstrate the superior performance of our algorithm.
****

Rectify Heterogeneous Models with Semantic Mapping
Han-Jia Ye, De-Chuan Zhan, Yuan Jiang, Zhi-Hua Zhou
On the way to the robust learner for real-world applications, there are still great challenges, including considering unknown environments with limited data. Learnware (Zhou; 2016) describes a novel perspective, and claims that learning models should have reusable and evolvable properties. We propose to Encode Meta InformaTion of features (EMIT), as the model specification for characterizing the changes, which grants the model evolvability to bridge heterogeneous feature spaces. Then, pre-trained models from related tasks can be Reused by our REctiFy via heterOgeneous pRedictor Mapping (REFORM}) framework. In summary, the pre-trained model is adapted to a new environment with different features, through model refining on only a small amount of training data in the current task. Experimental results over both synthetic and real-world tasks with diverse feature configurations validate the effectiveness and practical utility of the proposed framework.
****

Hierarchical Long-term Video Prediction without Supervision
Nevan wichers, Ruben Villegas, Dumitru Erhan, Honglak Lee
Much of recent research has been devoted to video prediction and generation, yet most of the previous works have demonstrated only limited success in generating videos on short-term horizons. The hierarchical video prediction method by Villegas et al. (2017) is an example of a state-of-the-art method for long-term video prediction, but their method is limited because it requires ground truth annotation of high-level structures (e.g., human joint landmarks) at training time. Our network encodes the input frame, predicts a high-level encoding into the future, and then a decoder with access to the first frame produces the predicted image from the predicted encoding. The decoder also produces a mask that outlines the predicted foreground object (e.g., person) as a by-product. Unlike Villegas et al. (2017), we develop a novel training method that jointly trains the encoder, the predictor, and the decoder together without highlevel supervision; we further improve upon this by using an adversarial loss in the feature space to train the predictor. Our method can predict about 20 seconds into the future and provides better results compared to Denton and Fergus (2018) and Finn et al. (2016) on the Human 3.6M dataset.
****