Making Algorithms Trustworthy: What Can Statistical Science Contribute to Transparency, Explanation and Validation?

Invited Talk (Breiman Lecture) | Thu Dec 6th 08:30 -- 09:20 AM @ Rooms 220 CDE

David Spiegelhalter

The demand for transparency, explainability and empirical validation of automated advice systems is not new. Back in the 1980s there were (occasionally acrimonious) discussions between proponents of rule-based systems and those based on statistical models, partly based on which were more transparent. A four-stage process of evaluation of medical advice systems was established, based on that used in drug development. More recently, EU legislation has focused attention on the ability of algorithms to, if required, show their workings. Inspired by Onora O'Neill's emphasis on demonstrating trustworthiness, and her idea of 'intelligent transparency', we should ideally be able to check (a) the empirical basis for the algorithm, (b) its past performance, (c) the reasoning behind its current claim, including tipping points and what-ifs (d) the uncertainty around its current claim, including whether the latest case comes within its remit. Furthermore, these explanations should be open to different levels of expertise. These ideas will be illustrated by the Predict 2.1 system for women choosing adjuvant therapy following surgery for breast cancer, which is based on a competing-risks survival regression model, and has been developed in collaboration with professional psychologists in close cooperation with clinicians and patients. Predict 2.1 has four levels of explanation of the claimed potential benefits and harms of alternative treatments, and is currently used in around 25,000 clinical decisions a month worldwide.

Coffee Break

Break | Thu Dec 6th 09:20 -- 09:45 AM @

—

Learning with SGD and Random Features

Spotlight | Thu Dec 6th 09:45 -- 09:50 AM @ Room 220 CD

Luigi Carratino · Alessandro Rudi · Lorenzo Rosasco

Sketching and stochastic gradient methods are arguably the most common techniques to derive efficient large scale learning algorithms. In this paper, we investigate their application in the context of nonparametric statistical learning. More precisely, we study the estimator defined by stochastic gradient with mini batches and random features. The latter can be seen as form of nonlinear sketching and used to define approximate kernel methods. The considered estimator is not explicitly penalized/constrained and regularization is implicit. Indeed, our study highlights how different parameters, such as number of features, iterations, step-size and mini-batch size control the learning properties of the solutions. We do this by deriving optimal finite sample bounds, under standard assumptions. The obtained results are corroborated and illustrated by numerical experiments.

Graphical model inference: Sequential Monte Carlo meets deterministic approximations

Spotlight | Thu Dec 6th 09:45 -- 09:50 AM @ Room 220 E

Fredrik Lindsten · Jouni Helske · Matti Vihola

Approximate inference in probabilistic graphical models (PGMs) can be grouped into deterministic methods and Monte-Carlo-based methods. The former can often provide accurate and rapid inferences, but are typically associated with biases that are hard to quantify. The latter enjoy asymptotic consistency, but can suffer from high computational costs. In this paper we present a way of bridging the gap between deterministic and stochastic inference. Specifically, we suggest an efficient sequential Monte Carlo (SMC) algorithm for PGMs which can leverage the output from deterministic inference methods. While generally applicable, we show explicitly how this can be done with loopy belief propagation, expectation propagation, and Laplace approximations. The resulting algorithm can be viewed as a post-correction of the biases associated with these methods and, indeed, numerical results show clear improvements over the baseline deterministic methods as well as over "plain" SMC.

Boolean Decision Rules via Column Generation

Spotlight | Thu Dec 6th 09:45 -- 09:50 AM @ Room 517 CD

Sanjeeb Dash · Oktay Gunluk · Dennis Wei

This paper considers the learning of Boolean rules in either disjunctive normal form (DNF, OR-of-ANDs, equivalent to decision rule sets) or conjunctive normal form (CNF, AND-of-ORs) as an interpretable model for classification. An integer program is formulated to optimally trade classification accuracy for rule simplicity. Column generation (CG) is used to efficiently search over an exponential number of candidate clauses (conjunctions or disjunctions) without the need for heuristic rule mining. This approach also bounds the gap between the selected rule set and the best possible rule set on the training data. To handle large datasets, we propose an approximate CG algorithm using randomization. Compared to three recently proposed alternatives, the CG algorithm dominates the accuracy-simplicity trade-off in 8 out of 16 datasets. When maximized for accuracy, CG is competitive with rule learners designed for this purpose, sometimes finding significantly simpler solutions that are no less accurate.

KONG: Kernels for ordered-neighborhood graphs

Spotlight | Thu Dec 6th 09:50 -- 09:55 AM @ Room 220 CD

Moez Draief · Konstantin Kutzkov · Kevin Scaman · Milan Vojnovic

We present novel graph kernels for graphs with node and edge labels that have ordered neighborhoods, i.e. when neighbor nodes follow an order. Graphs with ordered neighborhoods are a natural data representation for evolving graphs where edges are created over time, which induces an order. Combining convolutional subgraph kernels and string kernels, we design new scalable algorithms for generation of explicit graph feature maps using sketching techniques. We obtain precise bounds for the approximation accuracy and computational complexity of the proposed approaches and demonstrate their applicability on real datasets. In particular, our experiments demonstrate that neighborhood ordering results in more informative features. For the special case of general graphs, i.e. graphs without ordered neighborhoods, the new graph kernels yield efficient and simple algorithms for the comparison of label distributions between graphs.

Boosting Black Box Variational Inference

Spotlight | Thu Dec 6th 09:50 -- 09:55 AM @ Room 220 E

Francesco Locatello · Gideon Dresdner · Rajiv Khanna · Isabel Valera · Gunnar Raetsch

Approximating a probability density in a tractable manner is a central task in Bayesian statistics. Variational Inference (VI) is a popular technique that achieves tractability by choosing a relatively simple variational approximation. Borrowing ideas from the classic boosting framework, recent approaches attempt to \emph{boost} VI by replacing the selection of a single density with an iteratively constructed mixture of densities. In order to guarantee convergence, previous works impose stringent assumptions that require significant effort for practitioners. Specifically, they require a custom implementation of the greedy step (called the LMO) for every probabilistic model with respect to an unnatural variational family of truncated distributions. Our work fixes these issues with novel theoretical and algorithmic insights. On the theoretical side, we show that boosting VI satisfies a relaxed smoothness assumption which is sufficient for the convergence of the functional Frank-Wolfe (FW) algorithm. Furthermore, we rephrase the LMO problem and propose to maximize the Residual ELBO (RELBO) which replaces the standard ELBO optimization in VI. These theoretical enhancements allow for black box implementation of the boosting subroutine. Finally, we present a stopping criterion drawn from the duality gap in the classic FW analyses and exhaustive experiments to illustrate the usefulness of our theoretical and algorithmic contributions.

Fast greedy algorithms for dictionary selection with generalized sparsity constraints

Spotlight | Thu Dec 6th 09:50 -- 09:55 AM @ Room 517 CD

Kaito Fujii · Tasuku Soma

In dictionary selection, several atoms are selected from finite candidates that successfully approximate given data points in the sparse representation. We propose a novel efficient greedy algorithm for dictionary selection. Not only does our algorithm work much faster than the known methods, but it can also handle more complex sparsity constraints, such as average sparsity. Using numerical experiments, we show that our algorithm outperforms the known methods for dictionary selection, achieving competitive performances with dictionary learning algorithms in a smaller running time.

Quadrature-based features for kernel approximation

Spotlight | Thu Dec 6th 09:55 -- 10:00 AM @ Room 220 CD

Marina Munkhoeva · Yermek Kapushev · Evgeny Burnaev · Ivan Oseledets

We consider the problem of improving kernel approximation via randomized feature maps. These maps arise as Monte Carlo approximation to integral representations of kernel functions and scale up kernel methods for larger datasets. Based on an efficient numerical integration technique, we propose a unifying approach that reinterprets the previous random features methods and extends to better estimates of the kernel approximation. We derive the convergence behavior and conduct an extensive empirical study that supports our hypothesis.

Discretely Relaxing Continuous Variables for tractable Variational Inference

Spotlight | Thu Dec 6th 09:55 -- 10:00 AM @ Room 220 E

Trefor Evans · Prasanth Nair

We explore a new research direction in Bayesian variational inference with discrete latent variable priors where we exploit Kronecker matrix algebra for efficient and exact computations of the evidence lower bound (ELBO). The proposed "DIRECT" approach has several advantages over its predecessors; (i) it can exactly compute ELBO gradients (i.e. unbiased, zero-variance gradient estimates), eliminating the need for high-variance stochastic gradient estimators and enabling the use of quasi-Newton optimization methods; (ii) its training complexity is independent of the number of training points, permitting inference on large datasets; and (iii) its posterior samples consist of sparse and low-precision quantized integers which permit fast inference on hardware limited devices. In addition, our DIRECT models can exactly compute statistical moments of the parameterized predictive posterior without relying on Monte Carlo sampling. The DIRECT approach is not practical for all likelihoods, however, we identify a popular model structure which is practical, and demonstrate accurate inference using latent variables discretized as extremely low-precision 4-bit quantized integers. While the ELBO computations considered in the numerical studies require over 10^2352 log-likelihood evaluations, we train on datasets with over two-million points in just seconds.

Distributed k-Clustering for Data with Heavy Noise

Spotlight | Thu Dec 6th 09:55 -- 10:00 AM @ Room 517 CD

Shi Li · Xiangyu Guo

In this paper, we consider the k-center/median/means clustering with outliers problems (or the (k, z)-center/median/means problems) in the distributed setting. Most previous distributed algorithms have their communication costs linearly depending on z, the number of outliers. Recently Guha et al.[10] overcame this dependence issue by considering bi-criteria approximation algorithms that output solutions with $2z$ outliers. For the case where z is large, the extra z outliers discarded by the algorithms might be too large, considering that the data gathering process might be costly. In this paper, we improve the number of outliers to the best possible $(1+\epsilon)z$, while maintaining the $O(1)$-approximation ratio and independence of communication cost on z. The problems we consider include the (k, z)-center problem, and (k, z)-median/means problems in Euclidean metrics. Implementation of the our algorithm for (k, z)-center shows that it outperforms many previous algorithms, both in terms of the communication cost and quality of the output solution.

Statistical and Computational Trade-Offs in Kernel K-Means

Spotlight | Thu Dec 6th 10:00 -- 10:05 AM @ Room 220 CD

Daniele Calandriello · Lorenzo Rosasco

We investigate the efficiency of k-means in terms of both statistical and computational requirements. More precisely, we study a Nystr\"om approach to kernel k-means. We analyze the statistical properties of the proposed method and show that it achieves the same accuracy of exact kernel k-means with only a fraction of computations. Indeed, we prove under basic assumptions that sampling $\sqrt{n}$ Nystr\"om landmarks allows to greatly reduce computational costs without incurring in any loss of accuracy. To the best of our knowledge this is the first result showing in this kind for unsupervised learning.

Implicit Reparameterization Gradients

Spotlight | Thu Dec 6th 10:00 -- 10:05 AM @ Room 220 E

Mikhail Figurnov · Shakir Mohamed · Andriy Mnih

By providing a simple and efficient way of computing low-variance gradients of continuous random variables, the reparameterization trick has become the technique of choice for training a variety of latent variable models. However, it is not applicable to a number of important continuous distributions. We introduce an alternative approach to computing reparameterization gradients based on implicit differentiation and demonstrate its broader applicability by applying it to Gamma, Beta, Dirichlet, and von Mises distributions, which cannot be used with the classic reparameterization trick. Our experiments show that the proposed approach is faster and more accurate than the existing gradient estimators for these distributions.

Do Less, Get More: Streaming Submodular Maximization with Subsampling

Spotlight | Thu Dec 6th 10:00 -- 10:05 AM @ Room 517 CD

Moran Feldman · Amin Karbasi · Ehsan Kazemi

In this paper, we develop the first one-pass streaming algorithm for submodular maximization that does not evaluate the entire stream even once. By carefully subsampling each element of the data stream, our algorithm enjoys the tightest approximation guarantees in various settings while having the smallest memory footprint and requiring the lowest number of function evaluations. More specifically, for a monotone submodular function and a p-matchoid constraint, our randomized algorithm achieves a $4p$ approximation ratio (in expectation) with $O(k)$ memory and $O(km/p)$ queries per element (k is the size of the largest feasible solution and m is the number of matroids used to define the constraint). For the non-monotone case, our approximation ratio increases only slightly to $4p+2-o(1)$. To the best or our knowledge, our algorithm is the first that combines the benefits of streaming and subsampling in a novel way in order to truly scale submodular maximization to massive machine learning problems. To showcase its practicality, we empirically evaluated the performance of our algorithm on a video summarization application and observed that it outperforms the state-of-the-art algorithm by up to fifty-fold while maintaining practically the same utility. We also evaluated the scalability of our algorithm on a large dataset of Uber pick up locations.

Integrated accounts of behavioral and neuroimaging data using flexible recurrent neural network models

Oral | Thu Dec 6th 10:05 -- 10:20 AM @ Room 220 CD

Amir Dezfouli · Richard Morris · Fabio Ramos · Peter Dayan · Bernard Balleine

Neuroscience studies of human decision-making abilities commonly involve subjects completing a decision-making task while BOLD signals are recorded using fMRI. Hypotheses are tested about which brain regions mediate the effect of past experience, such as rewards, on future actions. One standard approach to this is model-based fMRI data analysis, in which a model is fitted to the behavioral data, i.e., a subject's choices, and then the neural data are parsed to find brain regions whose BOLD signals are related to the model's internal signals. However, the internal mechanics of such purely behavioral models are not constrained by the neural data, and therefore might miss or mischaracterize aspects of the brain. To address this limitation, we introduce a new method using recurrent neural network models that are flexible enough to be jointly fitted to the behavioral and neural data. We trained a model so that its internal states were suitably related to neural activity during the task, while at the same time its output predicted the next action a subject would execute. We then used the fitted model to create a novel visualization of the relationship between the activity in brain regions at different times following a reward and the choices the subject subsequently made. Finally, we validated our method using a previously published dataset. We found that the model was able to recover the underlying neural substrates that were discovered by explicit model engineering in the previous work, and also derived new results regarding the temporal pattern of brain activity.

Variational Inference with Tail-adaptive f-Divergence

Oral | Thu Dec 6th 10:05 -- 10:20 AM @ Room 220 E

Dilin Wang · Hao Liu · Qiang Liu

Variational inference with α-divergences has been widely used in modern probabilistic machine learning. Compared to Kullback-Leibler (KL) divergence, a major advantage of using α-divergences (with positive α values) is their mass-covering property. However, estimating and optimizing α-divergences require to use importance sampling, which could have extremely large or infinite variances due to heavy tails of importance weights. In this paper, we propose a new class of tail-adaptive f-divergences that adaptively change the convex function f with the tail of the importance weights, in a way that theoretically guarantee finite moments, while simultaneously achieving mass-covering properties. We test our methods on Bayesian neural networks, as well as deep reinforcement learning in which our method is applied to improve a recent soft actor-critic (SAC) algorithm (Haarnoja et al., 2018). Our results show that our approach yields significant advantages compared with existing methods based on classical KL and α-divergences.

Optimal Algorithms for Continuous Non-monotone Submodular and DR-Submodular Maximization

Oral | Thu Dec 6th 10:05 -- 10:20 AM @ Room 517 CD

Rad Niazadeh · Tim Roughgarden · Joshua Wang

In this paper we study the fundamental problems of maximizing a continuous non monotone submodular function over a hypercube, with and without coordinate-wise concavity. This family of optimization problems has several applications in machine learning, economics, and communication systems. Our main result is the first 1/2 approximation algorithm for continuous submodular function maximization; this approximation factor of is the best possible for algorithms that use only polynomially many queries. For the special case of DR-submodular maximization, we provide a faster 1/2-approximation algorithm that runs in (almost) linear time. Both of these results improve upon prior work [Bian et al., 2017, Soma and Yoshida, 2017, Buchbinder et al., 2012]. Our first algorithm is a single-pass algorithm that uses novel ideas such as reducing the guaranteed approximation problem to analyzing a zero-sum game for each coordinate, and incorporates the geometry of this zero-sum game to fix the value at this coordinate. Our second algorithm is a faster single-pass algorithm that exploits coordinate-wise concavity to identify a monotone equilibrium condition sufficient for getting the required approximation guarantee, and hunts for the equilibrium point using binary search. We further run experiments to verify the performance of our proposed algorithms in related machine learning applications.

Why Is My Classifier Discriminatory?

Spotlight | Thu Dec 6th 10:20 -- 10:25 AM @ Room 220 CD

Irene Chen · Fredrik Johansson · David Sontag

Recent attempts to achieve fairness in predictive models focus on the balance between fairness and accuracy. In sensitive applications such as healthcare or criminal justice, this trade-off is often undesirable as any increase in prediction error could have devastating consequences. In this work, we argue that the fairness of predictions should be evaluated in context of the data, and that unfairness induced by inadequate samples sizes or unmeasured predictive variables should be addressed through data collection, rather than by constraining the model. We decompose cost-based metrics of discrimination into bias, variance, and noise, and propose actions aimed at estimating and reducing each term. Finally, we perform case-studies on prediction of income, mortality, and review ratings, confirming the value of this analysis. We find that data collection is often a means to reduce discrimination without sacrificing accuracy.

Mirrored Langevin Dynamics

Spotlight | Thu Dec 6th 10:20 -- 10:25 AM @ Room 220 E

Ya-Ping Hsieh · Ali Kavis · Paul Rolland · Volkan Cevher

We consider the problem of sampling from constrained distributions, which has posed significant challenges to both non-asymptotic analysis and algorithmic design. We propose a unified framework, which is inspired by the classical mirror descent, to derive novel first-order sampling schemes. We prove that, for a general target distribution with strongly convex potential, our framework implies the existence of a first-order algorithm achieving O~(\epsilon^{-2}d) convergence, suggesting that the state-of-the-art O~(\epsilon^{-6}d^5) can be vastly improved. With the important Latent Dirichlet Allocation (LDA) application in mind, we specialize our algorithm to sample from Dirichlet posteriors, and derive the first non-asymptotic O~(\epsilon^{-2}d^2) rate for first-order sampling. We further extend our framework to the mini-batch setting and prove convergence rates when only stochastic gradients are available. Finally, we report promising experimental results for LDA on real datasets.

Overlapping Clustering Models, and One (class) SVM to Bind Them All

Spotlight | Thu Dec 6th 10:20 -- 10:25 AM @ Room 517 CD

Xueyu Mao · Purnamrita Sarkar · Deepayan Chakrabarti

People belong to multiple communities, words belong to multiple topics, and books cover multiple genres; overlapping clusters are commonplace. Many existing overlapping clustering methods model each person (or word, or book) as a non-negative weighted combination of "exemplars" who belong solely to one community, with some small noise. Geometrically, each person is a point on a cone whose corners are these exemplars. This basic form encompasses the widely used Mixed Membership Stochastic Blockmodel of networks and its degree-corrected variants, as well as topic models such as LDA. We show that a simple one-class SVM yields provably consistent parameter inference for all such models, and scales to large datasets. Experimental results on several simulated and real datasets show our algorithm (called SVM-cone) is both accurate and scalable.

Human-in-the-Loop Interpretability Prior

Spotlight | Thu Dec 6th 10:25 -- 10:30 AM @ Room 220 CD

Isaac Lage · Andrew Ross · Samuel J Gershman · Been Kim · Finale Doshi-Velez

We often desire our models to be interpretable as well as accurate. Prior work on optimizing models for interpretability has relied on easy-to-quantify proxies for interpretability, such as sparsity or the number of operations required. In this work, we optimize for interpretability by directly including humans in the optimization loop. We develop an algorithm that minimizes the number of user studies to find models that are both predictive and interpretable and demonstrate our approach on several data sets. Our human subjects results show trends towards different proxy notions of interpretability on different datasets, which suggests that different proxies are preferred on different tasks.

Global Convergence of Langevin Dynamics Based Algorithms for Nonconvex Optimization

Spotlight | Thu Dec 6th 10:25 -- 10:30 AM @ Room 220 E

Pan Xu · Jinghui Chen · Difan Zou · Quanquan Gu

We present a unified framework to analyze the global convergence of Langevin dynamics based algorithms for nonconvex finite-sum optimization with n component functions. At the core of our analysis is a direct analysis of the ergodicity of the numerical approximations to Langevin dynamics, which leads to faster convergence rates. Specifically, we show that gradient Langevin dynamics (GLD) and stochastic gradient Langevin dynamics (SGLD) converge to the \textit{almost minimizer}\footnote{Following \citet{raginsky2017non}, an almost minimizer is defined to be a point which is within the ball of the global minimizer with radius $O(d\log(\beta+1)/\beta)$, where d is the problem dimension and β is the inverse temperature parameter.} within $\tilde O\big(nd/(\lambda\epsilon) \big)$\footnote{$\tilde O(\cdot)$ notation hides polynomials of logarithmic terms and constants.} and $\tilde O\big(d^7/(\lambda^5\epsilon^5) \big)$ stochastic gradient evaluations respectively, where d is the problem dimension, and λ is the spectral gap of the Markov chain generated by GLD. Both results improve upon the best known gradient complexity\footnote{Gradient complexity is defined as the total number of stochastic gradient evaluations of an algorithm, which is the number of stochastic gradients calculated per iteration times the total number of iterations.} results \citep{raginsky2017non}. Furthermore, for the first time we prove the global convergence guarantee for variance reduced stochastic gradient Langevin dynamics (VR-SGLD) to the almost minimizer within $\tilde O\big(\sqrt{n}d^5/(\lambda^4\epsilon^{5/2})\big)$ stochastic gradient evaluations, which outperforms the gradient complexities of GLD and SGLD in a wide regime.
Our theoretical analyses shed some light on using Langevin dynamics based algorithms for nonconvex optimization with provable guarantees.

Removing the Feature Correlation Effect of Multiplicative Noise

Spotlight | Thu Dec 6th 10:25 -- 10:30 AM @ Room 517 CD

Zijun Zhang · Yining Zhang · Zongpeng Li

Multiplicative noise, including dropout, is widely used to regularize deep neural networks (DNNs), and is shown to be effective in a wide range of architectures and tasks. From an information perspective, we consider injecting multiplicative noise into a DNN as training the network to solve the task with noisy information pathways, which leads to the observation that multiplicative noise tends to increase the correlation between features, so as to increase the signal-to-noise ratio of information pathways. However, high feature correlation is undesirable, as it increases redundancy in representations. In this work, we propose non-correlating multiplicative noise (NCMN), which exploits batch normalization to remove the correlation effect in a simple yet effective way. We show that NCMN significantly improves the performance of standard multiplicative noise on image classification tasks, providing a better alternative to dropout for batch-normalized networks. Additionally, we present a unified view of NCMN and shake-shake regularization, which explains the performance gain of the latter.

Link Prediction Based on Graph Neural Networks

Spotlight | Thu Dec 6th 10:30 -- 10:35 AM @ Room 220 CD

Muhan Zhang · Yixin Chen

Link prediction is a key problem for network-structured data. Link prediction heuristics use some score functions, such as common neighbors and Katz index, to measure the likelihood of links. They have obtained wide practical uses due to their simplicity, interpretability, and for some of them, scalability. However, every heuristic has a strong assumption on when two nodes are likely to link, which limits their effectiveness on networks where these assumptions fail. In this regard, a more reasonable way should be learning a suitable heuristic from a given network instead of using predefined ones. By extracting a local subgraph around each target link, we aim to learn a function mapping the subgraph patterns to link existence, thus automatically learning a ``heuristic'' that suits the current network. In this paper, we study this heuristic learning paradigm for link prediction. First, we develop a novel γ-decaying heuristic theory. The theory unifies a wide range of heuristics in a single framework, and proves that all these heuristics can be well approximated from local subgraphs. Our results show that local subgraphs reserve rich information related to link existence. Second, based on the γ-decaying theory, we propose a new method to learn heuristics from local subgraphs using a graph neural network (GNN). Its experimental results show unprecedented performance, working consistently well on a wide range of problems.

Identification and Estimation of Causal Effects from Dependent Data

Spotlight | Thu Dec 6th 10:30 -- 10:35 AM @ Room 220 E

Eli Sherman · Ilya Shpitser

The assumption that data samples are independent and identically distributed (iid) is standard in many areas of statistics and machine learning. Nevertheless, in some settings, such as social networks, infectious disease modeling, and reasoning with spatial and temporal data, this assumption is false. An extensive literature exists on making causal inferences under the iid assumption [12, 8, 21, 16], but, as pointed out in [14], causal inference in non-iid contexts is challenging due to the combination of unobserved confounding bias and data dependence. In this paper we develop a general theory describing when causal inferences are possible in such scenarios. We use segregated graphs [15], a generalization of latent projection mixed graphs [23], to represent causal models of this type and provide a complete algorithm for non-parametric identification in these models. We then demonstrate how statistical inferences may be performed on causal parameters identified by this algorithm, even in cases where parts of the model exhibit full interference, meaning only a single sample is available for parts of the model [19]. We apply these techniques to a synthetic data set which considers the adoption of fake news articles given the social network structure, articles read by each person, and baseline demographics and socioeconomic covariates.

Connectionist Temporal Classification with Maximum Entropy Regularization

Spotlight | Thu Dec 6th 10:30 -- 10:35 AM @ Room 517 CD

Hu Liu · Sheng Jin · Changshui Zhang

Connectionist Temporal Classification (CTC) is an objective function for end-to-end sequence learning, which adopts dynamic programming algorithms to directly learn the mapping between sequences. CTC has shown promising results in many sequence learning applications including speech recognition and scene text recognition. However, CTC tends to produce highly peaky and overconfident distributions, which is a symptom of overfitting. To remedy this, we propose a regularization method based on maximum conditional entropy which penalizes peaky distributions and encourages exploration. We also introduce an entropy-based pruning method to dramatically reduce the number of CTC feasible paths by ruling out unreasonable alignments. Experiments on scene text recognition show that our proposed methods consistently improve over the CTC baseline without the need to adjust training settings. Code has been made publicly available at:

https://github.com/liuhu-bigeye/enctc.crnn.

Realistic Evaluation of Deep Semi-Supervised Learning Algorithms

Spotlight | Thu Dec 6th 10:35 -- 10:40 AM @ Room 220 CD

Avital Oliver · Augustus Odena · Colin A Raffel · Ekin Dogus Cubuk · Ian Goodfellow

Semi-supervised learning (SSL) provides a powerful framework for leveraging unlabeled data when labels are limited or expensive to obtain. SSL algorithms based on deep neural networks have recently proven successful on standard benchmark tasks. However, we argue that these benchmarks fail to address many issues that SSL algorithms would face in real-world applications. After creating a unified reimplementation of various widely-used SSL techniques, we test them in a suite of experiments designed to address these issues. We find that the performance of simple baselines which do not use unlabeled data is often underreported, SSL methods differ in sensitivity to the amount of labeled and unlabeled data, and performance can degrade substantially when the unlabeled dataset contains out-of-distribution examples. To help guide SSL research towards real-world applicability, we make our unified reimplemention and evaluation platform publicly available.

Causal Inference via Kernel Deviance Measures

Spotlight | Thu Dec 6th 10:35 -- 10:40 AM @ Room 220 E

Jovana Mitrovic · Dino Sejdinovic · Yee Whye Teh

Discovering the causal structure among a set of variables is a fundamental problem in many areas of science. In this paper, we propose Kernel Conditional Deviance for Causal Inference (KCDC) a fully nonparametric causal discovery method based on purely observational data. From a novel interpretation of the notion of asymmetry between cause and effect, we derive a corresponding asymmetry measure using the framework of reproducing kernel Hilbert spaces. Based on this, we propose three decision rules for causal discovery. We demonstrate the wide applicability and robustness of our method across a range of diverse synthetic datasets. Furthermore, we test our method on real-world time series data and the real-world benchmark dataset Tübingen Cause-Effect Pairs where we outperform state-of-the-art approaches.

Entropy and mutual information in models of deep neural

networks

Spotlight | Thu Dec 6th 10:35 -- 10:40 AM @ Room 517 CD

Marylou Gabrié · Andre Manoel · Clément Luneau · jean barbier · Nicolas Macris · Florent Krzakala · Lenka Zdeborová

We examine a class of stochastic deep learning models with a tractable method to compute information-theoretic quantities. Our contributions are three-fold: (i) We show how entropies and mutual informations can be derived from heuristic statistical physics methods, under the assumption that weight matrices are independent and orthogonally-invariant. (ii) We extend particular cases in which this result is known to be rigorously exact by providing a proof for two-layers networks with Gaussian random weights, using the recently introduced adaptive interpolation method. (iii) We propose an experiment framework with generative models of synthetic datasets, on which we train deep neural networks with a weight constraint designed so that the assumption in (i) is verified during learning. We study the behavior of entropies and mutual information throughout learning and conclude that, in the proposed setting, the relationship between compression and generalization remains elusive.

Automatic differentiation in ML: Where we are and where we should be going

Spotlight | Thu Dec 6th 10:40 -- 10:45 AM @ Room 220 CD

Bart van Merrienboer · Olivier Breuleux · Arnaud Bergeron · Pascal Lamblin

We review the current state of automatic differentiation (AD) for array programming in machine learning (ML), including the different approaches such as operator overloading (OO) and source transformation (ST) used for AD, graph-based intermediate representations for programs, and source languages. Based on these insights, we introduce a new graph-based intermediate representation (IR) which specifically aims to efficiently support fully-general AD for array programming. Unlike existing dataflow programming representations in ML frameworks, our IR naturally supports function calls, higher-order functions and recursion, making ML models easier to implement. The ability to represent closures allows us to perform AD using ST without a tape, making the resulting derivative (adjoint) program amenable to ahead-of-time optimization using tools from functional language compilers, and enabling higher-order derivatives. Lastly, we introduce a proof of concept compiler toolchain called Myia which uses a subset of Python as a front end.

Removing Hidden Confounding by Experimental Grounding

Spotlight | Thu Dec 6th 10:40 -- 10:45 AM @ Room 220 E

Nathan Kallus · Aahlad Manas Puli · Uri Shalit

Observational data is increasingly used as a means for making individual-level causal predictions and intervention recommendations. The foremost challenge of causal inference from observational data is hidden confounding, whose presence cannot be tested in data and can invalidate any causal conclusion. Experimental data does not suffer from confounding but is usually limited in both scope and scale. We introduce a novel method of using limited experimental data to correct the hidden confounding in causal effect models trained on larger observational data, even if the observational data does not fully overlap with the experimental data. Our method makes strictly weaker assumptions than existing approaches, and we prove conditions under which it yields a consistent estimator. We demonstrate our method's efficacy using real-world data from a large educational experiment.

The committee machine: Computational to statistical gaps in learning a two-layers neural network

Spotlight | Thu Dec 6th 10:40 -- 10:45 AM @ Room 517 CD

Benjamin Aubin · Antoine Maillard · jean barbier · Florent Krzakala · Nicolas Macris · Lenka Zdeborová

Heuristic tools from statistical physics have been used in the past to compute the optimal learning and generalization errors in the teacher-student scenario in multi- layer neural networks. In this contribution, we provide a rigorous justification of these approaches for a two-layers neural network model called the committee machine. We also introduce a version of the approximate message passing (AMP) algorithm for the committee machine that allows to perform optimal learning in polynomial time for a large set of parameters. We find that there are regimes in which a low generalization error is information-theoretically achievable while the AMP algorithm fails to deliver it; strongly suggesting that no efficient algorithm exists for those cases, and unveiling a large computational gap.

Experimental Design for Cost-Aware Learning of Causal Graphs

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #1

Erik Lindgren · Murat Kocaoglu · Alexandros Dimakis · Sriram Vishwanath

We consider the minimum cost intervention design problem: Given the essential graph of a causal graph and a cost to intervene on a variable, identify the set of interventions with minimum total cost that can learn any causal graph with the given essential graph. We first show that this problem is NP-hard. We then prove that we can achieve a constant factor approximation to this problem with a greedy algorithm. We then constrain the sparsity of each intervention. We develop an algorithm that returns an intervention design that is nearly optimal in terms of size for sparse graphs with sparse interventions and we discuss how to use it when there are costs on the vertices.

Removing Hidden Confounding by Experimental Grounding

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #2

Nathan Kallus · Aahlad Manas Puli · Uri Shalit

Observational data is increasingly used as a means for making individual-level causal predictions and intervention recommendations. The foremost challenge of causal inference from observational data is hidden confounding, whose presence cannot be tested in data and can invalidate any causal conclusion. Experimental data does not suffer from confounding but is usually limited in both scope and scale. We introduce a novel method of using limited experimental data to correct the hidden confounding in causal effect models trained on larger observational data, even if the observational data does not fully overlap with the experimental data. Our method makes strictly weaker assumptions than existing approaches, and we prove conditions under which it yields a consistent estimator. We demonstrate our method's efficacy using real-world data from a large educational experiment.

Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #3

Sara Magliacane · Thijs van Ommen · Tom Claassen · Stephan Bongers · Philip Versteeg · Joris M Mooij

An important goal common to domain adaptation and causal inference is to make accurate predictions when the distributions for the source (or training) domain(s) and target (or test) domain(s) differ. In many cases, these different distributions can be modeled as different contexts of a single underlying system, in which each distribution corresponds to a different perturbation of the system, or in causal terms, an intervention. We focus on a class of such causal domain adaptation problems, where data for one or more source domains are given, and the task is to predict the distribution of a certain target variable from measurements of other variables in one or more target domains. We propose an approach for solving these problems that exploits causal inference and does not rely on prior knowledge of the causal graph, the type of interventions or the intervention targets. We demonstrate our approach by evaluating a possible implementation on simulated and real world data.

Structural Causal Bandits: Where to Intervene?

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #4

Sanghack Lee · Elias Bareinboim

We study the problem of identifying the best action in a sequential decision-making setting when the reward distributions of the arms exhibit a non-trivial dependence structure, which is governed by the underlying causal model of the domain where the agent is deployed. In this setting, playing an arm corresponds to intervening on a set of variables and setting them to specific values. In this paper, we show that whenever the underlying causal model is not taken into account during the decision-making process, the standard strategies of simultaneously intervening on all variables or on all the subsets of the variables may, in general, lead to suboptimal policies, regardless of the number of interventions performed by the agent in the environment. We formally acknowledge this phenomenon and investigate structural properties implied by the underlying causal model, which lead to a complete characterization of the relationships between the arms' distributions. We leverage this characterization to build a new algorithm that takes as input a causal structure and finds a minimal, sound, and complete set of qualified arms that an agent should play to maximize its expected reward. We empirically demonstrate that the new strategy learns an optimal policy and leads to orders of magnitude faster convergence rates when compared with its causal-insensitive counterparts.

Uplift Modeling from Separate Labels

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #5

Ikko Yamane · Florian Yger · Jamal Atif · Masashi Sugiyama

Uplift modeling is aimed at estimating the incremental impact of an action on an individual's behavior, which is useful in various application domains such as targeted marketing (advertisement campaigns) and personalized medicine (medical treatments). Conventional methods of uplift modeling require every instance to be jointly equipped with two types of labels: the taken action and its outcome. However, obtaining two labels for each instance at the same time is difficult or expensive in many real-world problems. In this paper, we propose a novel method of uplift modeling that is applicable to a more practical setting where only one type of labels is available for each instance. We show a mean squared error bound for the proposed estimator and demonstrate its effectiveness through experiments.

Causal Inference with Noisy and Missing Covariates via Matrix Factorization

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #6

Nathan Kallus · Xiaojie Mao · Madeleine Udell

Valid causal inference in observational studies often requires controlling for confounders. However, in practice measurements of confounders may be noisy, and can lead to biased estimates of causal effects. We show that we can reduce bias induced by measurement noise using a large number of noisy measurements of the underlying confounders. We propose the use of matrix factorization to infer the confounders from noisy covariates. This flexible and principled framework adapts to missing values, accommodates a wide variety of data types, and can enhance a wide variety of causal inference methods. We bound the error for the induced average treatment effect estimator and show it is consistent in a linear regression setting, using Exponential Family Matrix Completion preprocessing. We demonstrate the effectiveness of the proposed procedure in numerical experiments with both synthetic data and real clinical data.

Fast Estimation of Causal Interactions using Wold Processes

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #7

Flavio Figueiredo · Guilherme Resende Borges · Pedro O.S. Vaz de Melo · Renato Assunção

We here focus on the task of learning Granger causality matrices for multivariate point processes. In order to accomplish this task, our work is the first to explore the use of Wold processes. By doing so, we are able to develop asymptotically fast MCMC learning algorithms. With N being the total number of events and K the number of processes, our learning algorithm has a $O(N(\,\log(N)\,+\,\log(K)))$ cost per iteration. This is much faster than the $O(N^3\,K^2)$ or $O(K^3)$ for the state of the art. Our approach, called GrangerBusca, is validated on nine datasets. This is an advance in relation to most prior efforts which focus mostly on subsets of the Memetracker data. Regarding accuracy, GrangerBusca is three times more accurate (in Precision@10) than the state of the art for the commonly explored subsets Memetracker. Due to GrangerBusca's much lower training complexity, our approach is the only one able to train models for larger, full, sets of data.

Learning and Testing Causal Models with Interventions

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #8

Jayadev Acharya · Arnab Bhattacharyya · Constantinos Daskalakis · Saravanan Kandasamy

We consider testing and learning problems on causal Bayesian networks as defined by Pearl (Pearl, 2009). Given a causal Bayesian network M on a graph with n discrete variables and bounded in-degree and bounded ``confounded components'', we show that O(log n) interventions on an unknown causal Bayesian network X on the same graph, and O(n/epsilon^2) samples per intervention, suffice to efficiently distinguish whether X=M or whether there exists some intervention under which X and M are farther than epsilon in total variation distance. We also obtain sample/time/intervention efficient algorithms for: (i) testing the identity of two unknown causal Bayesian networks on the same graph; and (ii) learning a causal Bayesian network on a given graph. Although our algorithms are non-adaptive, we show that adaptivity does not help in general: Omega(log n) interventions are necessary for testing the identity of two unknown causal Bayesian networks on the same graph, even adaptively. Our algorithms are enabled by a new subadditivity inequality for the squared Hellinger distance between two causal Bayesian networks.

Causal Inference via Kernel Deviance Measures

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #9

Jovana Mitrovic · Dino Sejdinovic · Yee Whye Teh

Discovering the causal structure among a set of variables is a fundamental problem in many areas of science. In this paper, we propose Kernel Conditional Deviance for Causal Inference (KCDC) a fully nonparametric causal discovery method based on purely observational data. From a novel interpretation of the notion of asymmetry between cause and effect, we derive a corresponding asymmetry measure using the framework of reproducing kernel Hilbert spaces. Based on this, we propose three decision rules for causal discovery. We demonstrate the wide applicability and robustness of our method across a range of diverse synthetic datasets. Furthermore, we test our method on real-world time series data and the real-world benchmark dataset Tübingen Cause-Effect Pairs where we outperform state-of-the-art approaches.

Multi-domain Causal Structure Learning in Linear Systems

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #10

AmirEmad Ghassami · Negar Kiyavash · Biwei Huang · Kun Zhang

We study the problem of causal structure learning in linear systems from observational data given in multiple domains, across which the causal coefficients and/or the distribution of the exogenous noises may vary. The main tool used in our approach is the principle that in a causally sufficient system, the causal modules, as well as their included parameters, change independently across domains. We first introduce our approach for finding causal direction in a system comprising two variables and propose efficient methods for identifying causal direction. Then we generalize our methods to causal structure learning in networks of variables. Most of previous work in structure learning from multi-domain data assume that certain types of invariance are held in causal modules across domains. Our approach unifies the idea in those works and generalizes to the case that there is no such invariance across the domains. Our proposed methods are generally capable of identifying causal direction from fewer than ten domains. When the invariance property holds, two domains are generally sufficient.

Causal Inference and Mechanism Clustering of A Mixture of Additive Noise Models

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #11

Shoubo Hu · Zhitang Chen · Vahid Partovi Nia · Laiwan CHAN · Yanhui Geng

The inference of the causal relationship between a pair of observed variables is a fundamental problem in science, and most existing approaches are based on one single causal model. In practice, however, observations are often collected from multiple sources with heterogeneous causal models due to certain uncontrollable factors, which renders causal analysis results obtained by a single model skeptical. In this paper, we generalize the Additive Noise Model (ANM) to a mixture model, which consists of a finite number of ANMs, and provide the condition of its causal identifiability. To conduct model estimation, we propose Gaussian Process Partially Observable Model (GPPOM), and incorporate independence enforcement into it to learn latent parameter associated with each observation. Causal inference and clustering according to the underlying generating mechanisms of the mixture model are addressed in this work. Experiments on synthetic and real data demonstrate the effectiveness of our proposed approach.

Direct Estimation of Differences in Causal Graphs

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #12

Yuhao Wang · Chandler Squires · Anastasiya Belyaeva · Caroline Uhler

We consider the problem of estimating the differences between two causal directed acyclic graph (DAG) models with a shared topological order given i.i.d. samples from each model. This is of interest for example in genomics, where changes in the structure or edge weights of the underlying causal graphs reflect alterations in the gene regulatory networks. We here provide the first provably consistent method for directly estimating the differences in a pair of causal DAGs without separately learning two possibly large and dense DAG models and computing their difference. Our two-step algorithm first uses invariance tests between regression coefficients of the two data sets to estimate the skeleton of the difference graph and then orients some of the edges using invariance tests between regression residual variances. We demonstrate the properties of our method through a simulation study and apply it to the analysis of gene expression data from ovarian cancer and during T-cell activation.

Identification and Estimation of Causal Effects from Dependent Data

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #13

Eli Sherman · Ilya Shpitser

The assumption that data samples are independent and identically distributed (iid) is standard in many areas of statistics and machine learning. Nevertheless, in some settings, such as social networks, infectious disease modeling, and reasoning with spatial and temporal data, this assumption is false. An extensive literature exists on making causal inferences under the iid assumption [12, 8, 21, 16], but, as pointed out in [14], causal inference in non-iid contexts is challenging due to the combination of unobserved confounding bias and data dependence. In this paper we develop a general theory describing when causal inferences are possible in such scenarios. We use segregated graphs [15], a generalization of latent projection mixed graphs [23], to represent causal models of this type and provide a complete algorithm for non-parametric identification in these models. We then demonstrate how statistical inferences may be performed on causal parameters identified by this algorithm, even in cases where parts of the model exhibit full interference, meaning only a single sample is available for parts of the model [19]. We apply these techniques to a synthetic data set which considers the adoption of fake news articles given the social network structure, articles read by each person, and baseline demographics and socioeconomic covariates.

Multilingual Anchoring: Interactive Topic Modeling and Alignment Across Languages

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #14

Michelle Yuan · Benjamin Van Durme · Jordan Ying

Multilingual topic models can reveal patterns in cross-lingual document collections. However, existing models lack speed and interactivity, which prevents adoption in everyday corpora exploration or quick moving situations (e.g., natural disasters, political instability). First, we propose a multilingual anchoring algorithm that builds an anchor-based topic model for documents in different languages. Then, we incorporate interactivity to develop MTAnchor (Multilingual Topic Anchors), a system that allows users to refine the topic model. We test our algorithms on labeled

English, Chinese, and Sinhalese documents. Within minutes, our methods can produce interpretable topics that are useful for specific classification tasks.

Submodular Field Grammars: Representation, Inference, and Application to Image Parsing

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #15

Abram L Friesen · Pedro Domingos

Natural scenes contain many layers of part-subpart structure, and distributions over them are thus naturally represented by stochastic image grammars, with one production per decomposition of a part. Unfortunately, in contrast to language grammars, where the number of possible split points for a production $A \rightarrow BC$ is linear in the length of A, in an image there are an exponential number of ways to split a region into subregions. This makes parsing intractable and requires image grammars to be severely restricted in practice, for example by allowing only rectangular regions. In this paper, we address this problem by associating with each production a submodular Markov random field whose labels are the subparts and whose labeling segments the current object into these subparts. We call the result a submodular field grammar (SFG). Finding the MAP split of a region into subregions is now tractable, and by exploiting this we develop an efficient approximate algorithm for MAP parsing of images with SFGs. Empirically, we present promising improvements in accuracy when using SFGs for scene understanding, and show exponential improvements in inference time compared to traditional methods, while returning comparable minima.

Autoconj: Recognizing and Exploiting Conjugacy Without a Domain-Specific Language

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #16

Matthew D. Hoffman · Matthew Johnson · Dustin Tran

Deriving conditional and marginal distributions using conjugacy relationships can be time consuming and error prone. In this paper, we propose a strategy for automating such derivations. Unlike previous systems which focus on relationships between pairs of random variables, our system (which we call Autoconj) operates directly on Python functions that compute log-joint distribution functions. Autoconj provides support for conjugacy-exploiting algorithms in any Python-embedded PPL. This paves the way for accelerating development of novel inference algorithms and structure-exploiting modeling strategies. The package can be downloaded at https://github.com/google-research/autoconj.

Distributionally Robust Graphical Models

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #17

Rizal Fathony · Ashkan Rezaei · Mohammad Ali Bashiri · Xinhua Zhang · Brian Ziebart

In many structured prediction problems, complex relationships between variables are compactly defined using graphical structures. The most prevalent graphical prediction methods---probabilistic graphical models and large margin methods---have their own distinct strengths but also possess significant drawbacks. Conditional random fields (CRFs) are Fisher consistent, but they do not permit integration of customized loss metrics into their learning process. Large-margin models, such as structured support vector machines (SSVMs), have the flexibility to incorporate customized loss metrics, but lack Fisher consistency guarantees. We present adversarial graphical models (AGM), a distributionally robust approach for constructing a predictor that performs robustly for a class of data distributions defined using a graphical structure. Our approach enjoys both the flexibility of incorporating customized loss metrics into its design as well as the statistical guarantee of Fisher consistency. We present exact learning and prediction algorithms for AGM with time complexity similar to existing graphical models and show the practical benefits of our approach with experiments.

Flexible and accurate inference and learning for deep generative models

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #18

Eszter Vértes · Maneesh Sahani

We introduce a new approach to learning in hierarchical latent-variable generative models called the “distributed distributional code Helmholtz machine”, which emphasises flexibility and accuracy in the inferential process. Like the original Helmholtz machine and later variational autoencoder algorithms (but unlike adver- sarial methods) our approach learns an explicit inference or “recognition” model to approximate the posterior distribution over the latent variables. Unlike these earlier methods, it employs a posterior representation that is not limited to a narrow tractable parametrised form (nor is it represented by samples). To train the genera- tive and recognition models we develop an extended wake-sleep algorithm inspired by the original Helmholtz machine. This makes it possible to learn hierarchical latent models with both discrete and continuous variables, where an accurate poste- rior representation is essential. We demonstrate that the new algorithm outperforms current state-of-the-art methods on synthetic, natural image patch and the MNIST data sets.

Provable Gaussian Embedding with One Observation

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #19

Ming Yu · Zhuoran Yang · Tuo Zhao · Mladen Kolar · Princeton Zhaoran Wang

The success of machine learning methods heavily relies on having an appropriate representation for data at hand. Traditionally, machine learning approaches relied on user-defined heuristics to extract features encoding structural information about data. However, recently there has been a surge in approaches that learn how to encode the data automatically in a low dimensional space. Exponential family embedding provides a probabilistic framework for learning low-dimensional representation for various types of high-dimensional data. Though successful in practice, theoretical underpinnings for exponential family embeddings have not been established. In this paper, we study the Gaussian embedding model and develop the first theoretical results for exponential family embedding models. First, we show that, under a mild condition, the embedding structure can be learned from one observation by leveraging the parameter sharing between different contexts even though the data are dependent with each other. Second, we study properties of two algorithms used for learning the embedding structure and establish convergence results for each of them. The first algorithm is based on a convex relaxation, while the other solved the non-convex formulation of the problem directly. Experiments demonstrate the effectiveness of our approach.

Learning and Inference in Hilbert Space with Quantum Graphical Models

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #20

Siddarth Srinivasan · Carlton Downey · Byron Boots

Quantum Graphical Models (QGMs) generalize classical graphical models by adopting the formalism for reasoning about uncertainty from quantum mechanics. Unlike classical graphical models, QGMs represent uncertainty with density matrices in complex Hilbert spaces. Hilbert space embeddings (HSEs) also generalize Bayesian inference in Hilbert spaces. We investigate the link between QGMs and HSEs and show that the sum rule and Bayes rule for QGMs are equivalent to the kernel sum rule in HSEs and a special case of Nadaraya-Watson kernel regression, respectively. We show that these operations can be kernelized, and use these insights to propose a Hilbert Space Embedding of Hidden Quantum Markov Models (HSE-HQMM) to model dynamics. We present experimental results showing that HSE-HQMMs are competitive with state-of-the-art models like LSTMs and PSRNNs on several datasets, while also providing a nonparametric method for maintaining a probability distribution over continuous-valued features.

Multi-value Rule Sets for Interpretable Classification with Feature-Efficient Representations

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #21

Tong Wang

We present the Multi-value Rule Set (MRS) for interpretable classification with feature efficient presentations. Compared to rule sets built from single-value rules, MRS adopts a more generalized form of association rules that allows multiple values in a condition. Rules of this form are more concise than classical single-value rules in capturing and describing patterns in data. Our formulation also pursues a higher efficiency of feature utilization, which reduces possible cost in data collection and storage. We propose a Bayesian framework for formulating an MRS model and develop an efficient inference method for learning a maximum a posteriori, incorporating theoretically grounded bounds to iteratively reduce the search space and improve the search efficiency. Experiments on synthetic and real-world data demonstrate that MRS models have significantly smaller complexity and fewer features than baseline models while being competitive in predictive accuracy.

Nonparametric Bayesian Lomax delegate racing for survival analysis with competing risks

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #22

Quan Zhang · Mingyuan Zhou

We propose Lomax delegate racing (LDR) to explicitly model the mechanism of survival under competing risks and to interpret how the covariates accelerate or decelerate the time to event. LDR explains non-monotonic covariate effects by racing a potentially infinite number of sub-risks, and consequently relaxes the ubiquitous proportional-hazards assumption which may be too restrictive. Moreover, LDR is naturally able to model not only censoring, but also missing event times or event types. For inference, we develop a Gibbs sampler under data augmentation for moderately sized data, along with a stochastic gradient descent maximum a posteriori inference algorithm for big data applications. Illustrative experiments are provided on both synthetic and real datasets, and comparison with various benchmark algorithms for survival analysis with competing risks demonstrates distinguished performance of LDR.

Theoretical guarantees for EM under misspecified Gaussian mixture models

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #23

Raaz Dwivedi · nhật Hồ · Koulik Khamaru · Martin Wainwright · Michael Jordan

Recent years have witnessed substantial progress in understanding the behavior of EM for mixture models that are correctly specified. Given that model misspecification is common in practice, it is important to understand EM in this more general setting. We provide non-asymptotic guarantees for population and sample-based EM for parameter estimation under a few specific univariate settings of misspecified Gaussian mixture models. Due to misspecification, the EM iterates no longer converge to the true model and instead converge to the projection of the true model over the set of models being searched over. We provide two classes of theoretical guarantees: first, we characterize the bias introduced due to the misspecification; and second, we prove that population EM converges at a geometric rate to the model projection under a suitable initialization condition. This geometric convergence rate for population EM imply a statistical complexity of order $1/\sqrt{n}$ when running EM with n samples. We validate our theoretical findings in different cases via several numerical examples.

Nonparametric learning from Bayesian models with randomized objective functions

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #24

Simon Lyddon · Stephen Walker · Chris C Holmes

Bayesian learning is built on an assumption that the model space contains a true reflection of the data generating mechanism. This assumption is problematic, particularly in complex data environments. Here we present a Bayesian nonparametric approach to learning that makes use of statistical models, but does not assume that the model is true. Our approach has provably better properties than using a parametric model and admits a Monte Carlo sampling scheme that can afford massive scalability on modern computer architectures. The model-based aspect of learning is particularly attractive for regularizing nonparametric inference when the sample size is small, and also for correcting approximate approaches such as variational Bayes (VB). We demonstrate the approach on a number of examples including VB classifiers and Bayesian random forests.

Rectangular Bounding Process

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #25

Xuhui Fan · Bin Li · Scott SIsson

Stochastic partition models divide a multi-dimensional space into a number of rectangular regions, such that the data within each region exhibit certain types of homogeneity. Due to the nature of their partition strategy, existing partition models may create many unnecessary divisions in sparse regions when trying to describe data in dense regions. To avoid this problem we introduce a new parsimonious partition model -- the Rectangular Bounding Process (RBP) -- to efficiently partition multi-dimensional spaces, by employing a bounding strategy to enclose data points within rectangular bounding boxes. Unlike existing approaches, the RBP possesses several attractive theoretical properties that make it a powerful nonparametric partition prior on a hypercube. In particular, the RBP is self-consistent and as such can be directly extended from a finite hypercube to infinite (unbounded) space. We apply the RBP to regression trees and relational models as a flexible partition prior. The experimental results validate the merit of the RBP {in rich yet parsimonious expressiveness} compared to the state-of-the-art methods.

A Bayesian Nonparametric View on Count-Min Sketch

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #26

Diana Cai · Michael Mitzenmacher · Ryan Adams

The count-min sketch is a time- and memory-efficient randomized data structure that provides a point estimate of the number of times an item has appeared in a data stream. The count-min sketch and related hash-based data structures are ubiquitous in systems that must track frequencies of data such as URLs, IP addresses, and language n-grams. We present a Bayesian view on the count-min sketch, using the same data structure, but providing a posterior distribution over the frequencies that characterizes the uncertainty arising from the hash-based approximation. In particular, we take a nonparametric approach and consider tokens generated from a Dirichlet process (DP) random measure, which allows for an unbounded number of unique tokens. Using properties of the DP, we show that it is possible to straightforwardly compute posterior marginals of the unknown true counts and that the modes of these marginals recover the count-min sketch estimator, inheriting the associated probabilistic guarantees. Using simulated data with known ground truth, we investigate the properties of these estimators. Lastly, we also study a modified problem in which the observation stream consists of collections of tokens (i.e., documents) arising from a random measure drawn from a stable beta process, which allows for power law scaling behavior in the number of unique tokens.

Communication Efficient Parallel Algorithms for Optimization on Manifolds

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #27

Bayan Saparbayeva · Michael Zhang · Lizhen Lin

The last decade has witnessed an explosion in the development of models, theory and computational algorithms for ``big data'' analysis. In particular, distributed inference has served as a natural and dominating paradigm for statistical inference. However, the existing literature on parallel inference almost exclusively focuses on Euclidean data and parameters. While this assumption is valid for many applications, it is increasingly more common to encounter problems where the data or the parameters lie on a non-Euclidean space, like a manifold for example. Our work aims to fill a critical gap in the literature by generalizing parallel inference algorithms to optimization on manifolds. We show that our proposed algorithm is both communication efficient and carries theoretical convergence guarantees. In addition, we demonstrate the performance of our algorithm to the estimation of Fr\'echet means on simulated spherical data and the low-rank matrix completion problem over Grassmann manifolds applied to the Netflix prize data set.

Lifted Weighted Mini-Bucket

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #28

Nicholas Gallo · Alexander Ihler

Many graphical models, such as Markov Logic Networks (MLNs) with evidence, possess highly symmetric substructures but no exact symmetries. Unfortunately, there are few principled methods that exploit these symmetric substructures to perform efficient approximate inference. In this paper, we present a lifted variant of the Weighted Mini-Bucket elimination algorithm which provides a principled way to (i) exploit the highly symmetric substructure of MLN models, and (ii) incorporate high-order inference terms which are necessary for high quality approximate inference. Our method has significant control over the accuracy-time trade-off of the approximation, allowing us to generate any-time approximations. Experimental results demonstrate the utility of this class of approximations, especially in models with strong repulsive potentials.

Cluster Variational Approximations for Structure Learning of Continuous-Time Bayesian Networks from Incomplete Data

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #29

Dominik Linzner · Heinz Koeppl

Continuous-time Bayesian networks (CTBNs) constitute a general and powerful framework for modeling continuous-time stochastic processes on networks. This makes them particularly attractive for learning the directed structures among interacting entities. However, if the available data is incomplete, one needs to simulate the prohibitively complex CTBN dynamics. Existing approximation techniques, such as sampling and low-order variational methods, either scale unfavorably in system size, or are unsatisfactory in terms of accuracy. Inspired by recent advances in statistical physics, we present a new approximation scheme based on cluster-variational methods that significantly improves upon existing variational approximations. We can analytically marginalize the parameters of the approximate CTBN, as these are of secondary importance for structure learning. This recovers a scalable scheme for direct structure learning from incomplete and noisy time-series data. Our approach outperforms existing methods in terms of scalability.

Faithful Inversion of Generative Models for Effective Amortized Inference

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #30

Stefan Webb · Adam Golinski · Rob Zinkov · Siddharth Narayanaswamy · Tom Rainforth · Yee Whye Teh · Frank Wood

Inference amortization methods share information across multiple posterior-inference problems, allowing each to be carried out more efficiently. Generally, they require the inversion of the dependency structure in the generative model, as the modeller must learn a mapping from observations to distributions approximating the posterior. Previous approaches have involved inverting the dependency structure in a heuristic way that fails to capture these dependencies correctly, thereby limiting the achievable accuracy of the resulting approximations. We introduce an algorithm for faithfully, and minimally, inverting the graphical model structure of any generative model. Such inverses have two crucial properties: (a) they do not encode any independence assertions that are absent from the model and; (b) they are local maxima for the number of true independencies encoded. We prove the correctness of our approach and empirically show that the resulting minimally faithful inverses lead to better inference amortization than existing heuristic approaches.

A Stein variational Newton method

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #31

Gianluca Detommaso · Tiangang Cui · Youssef Marzouk · Alessio Spantini · Robert Scheichl

Stein variational gradient descent (SVGD) was recently proposed as a general purpose nonparametric variational inference algorithm: it minimizes the Kullback–Leibler divergence between the target distribution and its approximation by implementing a form of functional gradient descent on a reproducing kernel Hilbert space [Liu & Wang, NIPS 2016]. In this paper, we accelerate and generalize the SVGD algorithm by including second-order information, thereby approximating a Newton-like iteration in function space. We also show how second-order information can lead to more effective choices of kernel. We observe significant computational gains over the original SVGD algorithm in multiple test cases.

Reparameterization Gradient for Non-differentiable Models

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #32

Wonyeol Lee · Hangyeol Yu · Hongseok Yang

We present a new algorithm for stochastic variational inference that targets at models with non-differentiable densities. One of the key challenges in stochastic variational inference is to come up with a low-variance estimator of the gradient of a variational objective. We tackle the challenge by generalizing the reparameterization trick, one of the most effective techniques for addressing the variance issue for differentiable models, so that the trick works for non-differentiable models as well. Our algorithm splits the space of latent variables into regions where the density of the variables is differentiable, and their boundaries where the density may fail to be differentiable. For each differentiable region, the algorithm applies the standard reparameterization trick and estimates the gradient restricted to the region. For each potentially non-differentiable boundary, it uses a form of manifold sampling and computes the direction for variational parameters that, if followed, would increase the boundary’s contribution to the variational objective. The sum of all the estimates becomes the gradient estimate of our algorithm. Our estimator enjoys the reduced variance of the reparameterization gradient while remaining unbiased even for non-differentiable models. The experiments with our preliminary implementation confirm the benefit of reduced variance and unbiasedness.

Implicit Reparameterization Gradients

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #33

Mikhail Figurnov · Shakir Mohamed · Andriy Mnih

By providing a simple and efficient way of computing low-variance gradients of continuous random variables, the reparameterization trick has become the technique of choice for training a variety of latent variable models. However, it is not applicable to a number of important continuous distributions. We introduce an alternative approach to computing reparameterization gradients based on implicit differentiation and demonstrate its broader applicability by applying it to Gamma, Beta, Dirichlet, and von Mises distributions, which cannot be used with the classic reparameterization trick. Our experiments show that the proposed approach is faster and more accurate than the existing gradient estimators for these distributions.

SLANG: Fast Structured Covariance Approximations for Bayesian Deep Learning with Natural Gradient

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #34

Aaron Mishkin · Frederik Kunstner · Didrik Nielsen · Mark Schmidt · Mohammad Emtiyaz Khan

Uncertainty estimation in large deep-learning models is a computationally challenging task, where it is difficult to form even a Gaussian approximation to the posterior distribution. In such situations, existing methods usually resort to a diagonal approximation of the covariance matrix despite the fact that these matrices are known to give poor uncertainty estimates. To address this issue, we propose a new stochastic, low-rank, approximate natural-gradient (SLANG) method for variational inference in large deep models. Our method estimates a “diagonal plus low-rank” structure based solely on back-propagated gradients of the network log-likelihood. This requires strictly less gradient computations than methods that compute the gradient of the whole variational objective. Empirical evaluations on standard benchmarks confirm that SLANG enables faster and more accurate estimation of uncertainty than mean-field methods, and performs comparably to state-of-the-art methods.

Wasserstein Variational Inference

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #35

Luca Ambrogioni · Umut Güçlü · Yağmur Güçlütürk · Max Hinne · Marcel A. J. van Gerven · Eric Maris

This paper introduces Wasserstein variational inference, a new form of approximate Bayesian inference based on optimal transport theory. Wasserstein variational inference uses a new family of divergences that includes both f-divergences and the Wasserstein distance as special cases. The gradients of the Wasserstein variational loss are obtained by backpropagating through the Sinkhorn iterations. This technique results in a very stable likelihood-free training method that can be used with implicit distributions and probabilistic programs. Using the Wasserstein variational inference framework, we introduce several new forms of autoencoders and test their robustness and performance against existing variational autoencoding techniques.

Adaptive Path-Integral Autoencoders: Representation

Learning and Planning for Dynamical Systems

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #36

Jung-Su Ha · Young-Jin Park · Hyeok-Joo Chae · Soon-Seo Park · Han-Lim Choi

We present a representation learning algorithm that learns a low-dimensional latent dynamical system from high-dimensional sequential raw data, e.g., video. The framework builds upon recent advances in amortized inference methods that use both an inference network and a refinement procedure to output samples from a variational distribution given an observation sequence, and takes advantage of the duality between control and inference to approximately solve the intractable inference problem using the path integral control approach. The learned dynamical model can be used to predict and plan the future states; we also present the efficient planning method that exploits the learned low-dimensional latent dynamics. Numerical experiments show that the proposed path-integral control based variational inference method leads to tighter lower bounds in statistical model learning of sequential data. Supplementary video: https://youtu.be/xCp35crUoLQ

Variational Inference with Tail-adaptive f-Divergence

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #37

Dilin Wang · Hao Liu · Qiang Liu

Variational inference with α-divergences has been widely used in modern probabilistic machine learning. Compared to Kullback-Leibler (KL) divergence, a major advantage of using α-divergences (with positive α values) is their mass-covering property. However, estimating and optimizing α-divergences require to use importance sampling, which could have extremely large or infinite variances due to heavy tails of importance weights. In this paper, we propose a new class of tail-adaptive f-divergences that adaptively change the convex function f with the tail of the importance weights, in a way that theoretically guarantee finite moments, while simultaneously achieving mass-covering properties. We test our methods on Bayesian neural networks, as well as deep reinforcement learning in which our method is applied to improve a recent soft actor-critic (SAC) algorithm (Haarnoja et al., 2018). Our results show that our approach yields significant advantages compared with existing methods based on classical KL and α-divergences.

Boosting Black Box Variational Inference

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #38

Francesco Locatello · Gideon Dresdner · Rajiv Khanna · Isabel Valera · Gunnar Raetsch

Approximating a probability density in a tractable manner is a central task in Bayesian statistics.

Variational Inference (VI) is a popular technique that achieves tractability by choosing a relatively simple variational approximation. Borrowing ideas from the classic boosting framework, recent approaches attempt to \emph{boost} VI by replacing the selection of a single density with an iteratively constructed mixture of densities. In order to guarantee convergence, previous works impose stringent assumptions that require significant effort for practitioners. Specifically, they require a custom implementation of the greedy step (called the LMO) for every probabilistic model with respect to an unnatural variational family of truncated distributions. Our work fixes these issues with novel theoretical and algorithmic insights. On the theoretical side, we show that boosting VI satisfies a relaxed smoothness assumption which is sufficient for the convergence of the functional Frank-Wolfe (FW) algorithm. Furthermore, we rephrase the LMO problem and propose to maximize the Residual ELBO (RELBO) which replaces the standard ELBO optimization in VI. These theoretical enhancements allow for black box implementation of the boosting subroutine. Finally, we present a stopping criterion drawn from the duality gap in the classic FW analyses and exhaustive experiments to illustrate the usefulness of our theoretical and algorithmic contributions.

Discretely Relaxing Continuous Variables for tractable Variational Inference

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #39

Trefor Evans · Prasanth Nair

We explore a new research direction in Bayesian variational inference with discrete latent variable priors where we exploit Kronecker matrix algebra for efficient and exact computations of the evidence lower bound (ELBO). The proposed "DIRECT" approach has several advantages over its predecessors; (i) it can exactly compute ELBO gradients (i.e. unbiased, zero-variance gradient estimates), eliminating the need for high-variance stochastic gradient estimators and enabling the use of quasi-Newton optimization methods; (ii) its training complexity is independent of the number of training points, permitting inference on large datasets; and (iii) its posterior samples consist of sparse and low-precision quantized integers which permit fast inference on hardware limited devices. In addition, our DIRECT models can exactly compute statistical moments of the parameterized predictive posterior without relying on Monte Carlo sampling. The DIRECT approach is not practical for all likelihoods, however, we identify a popular model structure which is practical, and demonstrate accurate inference using latent variables discretized as extremely low-precision 4-bit quantized integers. While the ELBO computations considered in the numerical studies require over 10^2352 log-likelihood evaluations, we train on datasets with over two-million points in just seconds.

Using Large Ensembles of Control Variates for Variational Inference

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #40

Tomas Geffner · Justin Domke

Variational inference is increasingly being addressed with stochastic optimization. In this setting, the gradient's variance plays a crucial role in the optimization procedure, since high variance gradients lead to poor convergence. A popular approach used to reduce gradient's variance involves the use of control variates. Despite the good results obtained, control variates developed for variational inference are typically looked at in isolation. In this paper we clarify the large number of control variates that are available by giving a systematic view of how they are derived. We also present a Bayesian risk minimization framework in which the quality of a procedure for combining control variates is quantified by its effect on optimization convergence rates, which leads to a very simple combination rule. Results show that combining a large number of control variates this way significantly improves the convergence of inference over using the typical gradient estimators or a reduced number of control variates.

The promises and pitfalls of Stochastic Gradient Langevin Dynamics

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #41

Nicolas Brosse · Alain Durmus · Eric Moulines

Stochastic Gradient Langevin Dynamics (SGLD) has emerged as a key MCMC algorithm for Bayesian learning from large scale datasets. While SGLD with decreasing step sizes converges weakly to the posterior distribution, the algorithm is often used with a constant step size in practice and has demonstrated spectacular successes in machine learning tasks. The current practice is to set the step size inversely proportional to N where N is the number of training samples. As N becomes large, we show that the SGLD algorithm has an invariant probability measure which significantly departs from the target posterior and behaves like as Stochastic Gradient Descent (SGD). This difference is inherently due to the high variance of the stochastic gradients. Several strategies have been suggested to reduce this effect; among them, SGLD Fixed Point (SGLDFP) uses carefully designed control variates to reduce the variance of the stochastic gradients. We show that SGLDFP gives approximate samples from the posterior distribution, with an accuracy comparable to the Langevin Monte Carlo (LMC) algorithm for a computational cost sublinear in the number of data points. We provide a detailed analysis of the Wasserstein distances between LMC, SGLD, SGLDFP and SGD and explicit expressions of the means and covariance matrices of their invariant distributions. Our findings are supported by limited numerical experiments.

Large-Scale Stochastic Sampling from the Probability Simplex

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #42

Jack Baker · Paul Fearnhead · Emily Fox · Christopher Nemeth

Stochastic gradient Markov chain Monte Carlo (SGMCMC) has become a popular method for scalable Bayesian inference. These methods are based on sampling a discrete-time approximation to a continuous time process, such as the Langevin diffusion. When applied to distributions defined on a constrained space the time-discretization error can dominate when we are near the boundary of the space. We demonstrate that because of this, current SGMCMC methods for the simplex struggle with sparse simplex spaces; when many of the components are close to zero. Unfortunately, many popular large-scale Bayesian models, such as network or topic models, require inference on sparse simplex spaces. To avoid the biases caused by this discretization error, we propose the stochastic Cox-Ingersoll-Ross process (SCIR), which removes all discretization error and we prove that samples from the SCIR process are asymptotically unbiased. We discuss how this idea can be extended to target other constrained spaces. Use of the SCIR process within a SGMCMC algorithm is shown to give substantially better performance for a topic model and a Dirichlet process mixture model than existing SGMCMC approaches.

Mirrored Langevin Dynamics

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #43

Ya-Ping Hsieh · Ali Kavis · Paul Rolland · Volkan Cevher

We consider the problem of sampling from constrained distributions, which has posed significant challenges to both non-asymptotic analysis and algorithmic design. We propose a unified framework, which is inspired by the classical mirror descent, to derive novel first-order sampling schemes. We prove that, for a general target distribution with strongly convex potential, our framework implies the existence of a first-order algorithm achieving O~(\epsilon^{-2}d) convergence, suggesting that the state-of-the-art O~(\epsilon^{-6}d^5) can be vastly improved. With the important Latent Dirichlet Allocation (LDA) application in mind, we specialize our algorithm to sample from Dirichlet posteriors, and derive the first non-asymptotic O~(\epsilon^{-2}d^2) rate for first-order sampling. We further extend our framework to the mini-batch setting and prove convergence rates when only stochastic gradients are available. Finally, we report promising experimental results for LDA on real datasets.

Thermostat-assisted continuously-tempered Hamiltonian Monte Carlo for Bayesian learning

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #44

Rui Luo · Jianhong Wang · Yaodong Yang · Jun WANG · Zhanxing Zhu

In this paper, we propose a novel sampling method, the thermostat-assisted continuously-tempered Hamiltonian Monte Carlo, for the purpose of multimodal Bayesian learning. It simulates a noisy dynamical system by incorporating both a continuously-varying tempering variable and the Nos\'e-Hoover thermostats. A significant benefit is that it is not only able to efficiently generate i.i.d. samples when the underlying posterior distributions are multimodal, but also capable of adaptively neutralising the noise arising from the use of mini-batches. While the properties of the approach have been studied using synthetic datasets, our experiments on three real datasets have also shown its performance gains over several strong baselines for Bayesian learning with various types of neural networks plunged in.

Dimensionally Tight Bounds for Second-Order Hamiltonian Monte Carlo

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #45

Oren Mangoubi · Nisheeth Vishnoi

Hamiltonian Monte Carlo (HMC) is a widely deployed method to sample from high-dimensional distributions in Statistics and Machine learning. HMC is known to run very efficiently in practice and its popular second-order `leapfrog" implementation has long been conjectured to run in $d^{1/4}$ gradient evaluations. Here we show that this conjecture is true when sampling from strongly log-concave target distributions that satisfy a weak third-order regularity property associated with the input data. Our regularity condition is weaker than the Lipschitz Hessian property and allows us to show faster convergence bounds for a much larger class of distributions than would be possible with the usual Lipschitz Hessian constant alone. Important distributions that satisfy our regularity condition include posterior distributions used in Bayesian logistic regression for which the data satisfies an`incoherence" property. Our result compares favorably with the best available bounds for the class of strongly log-concave distributions, which grow like $d^{{1}/{2}}$ gradient evaluations with the dimension. Moreover, our simulations on synthetic data suggest that, when our regularity condition is satisfied, leapfrog HMC performs better than its competitors -- both in terms of accuracy and in terms of the number of gradient evaluations it requires.

Global Convergence of Langevin Dynamics Based Algorithms for Nonconvex Optimization

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #46

Pan Xu · Jinghui Chen · Difan Zou · Quanquan Gu

We present a unified framework to analyze the global convergence of Langevin dynamics based algorithms for nonconvex finite-sum optimization with n component functions. At the core of our analysis is a direct analysis of the ergodicity of the numerical approximations to Langevin dynamics, which leads to faster convergence rates. Specifically, we show that gradient Langevin dynamics (GLD) and stochastic gradient Langevin dynamics (SGLD) converge to the \textit{almost minimizer}\footnote{Following \citet{raginsky2017non}, an almost minimizer is defined to be a point which is within the ball of the global minimizer with radius $O(d\log(\beta+1)/\beta)$, where d is the problem dimension and β is the inverse temperature parameter.} within $\tilde O\big(nd/(\lambda\epsilon) \big)$\footnote{$\tilde O(\cdot)$ notation hides polynomials of logarithmic terms and constants.} and $\tilde O\big(d^7/(\lambda^5\epsilon^5) \big)$ stochastic gradient evaluations respectively, where d is the problem dimension, and λ is the spectral gap of the Markov chain generated by GLD. Both results improve upon the best known gradient complexity\footnote{Gradient complexity is defined as the total number of stochastic gradient evaluations of an algorithm, which is the number of stochastic gradients calculated per iteration times the total number of iterations.} results \citep{raginsky2017non}. Furthermore, for the first time we prove the global convergence guarantee for variance reduced stochastic gradient Langevin dynamics (VR-SGLD) to the almost minimizer within $\tilde O\big(\sqrt{n}d^5/(\lambda^4\epsilon^{5/2})\big)$ stochastic gradient evaluations, which outperforms the gradient complexities of GLD and SGLD in a wide regime.
Our theoretical analyses shed some light on using Langevin dynamics based algorithms for nonconvex optimization with provable guarantees.

Meta-Learning MCMC Proposals

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #47

Tongzhou Wang · YI WU · Dave Moore · Stuart Russell

Effective implementations of sampling-based probabilistic inference often require manually constructed, model-specific proposals. Inspired by recent progresses in meta-learning for training learning agents that can generalize to unseen environments, we propose a meta-learning approach to building effective and generalizable MCMC proposals. We parametrize the proposal as a neural network to provide fast approximations to block Gibbs conditionals. The learned neural proposals generalize to occurrences of common structural motifs across different models, allowing for the construction of a library of learned inference primitives that can accelerate inference on unseen

models with no model-specific training required. We explore several applications including open-universe Gaussian mixture models, in which our learned proposals outperform a hand-tuned sampler, and a real-world named entity recognition task, in which our sampler yields higher final F1 scores than classical single-site Gibbs sampling.

Posterior Concentration for Sparse Deep Learning

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #48

Veronika Rockova · nicholas polson

We introduce Spike-and-Slab Deep Learning (SS-DL), a fully Bayesian alternative to dropout for improving generalizability of deep ReLU networks. This new type of regularization enables provable recovery of smooth input-output maps with {\sl unknown} levels of smoothness. Indeed, we show that the posterior distribution concentrates at the near minimax rate for alpha-Holder smooth maps, performing as well as if we knew the smoothness level alpha ahead of time. Our result sheds light on architecture design for deep neural networks, namely the choice of depth, width and sparsity level. These network attributes typically depend on unknown smoothness in order to be optimal. We obviate this constraint with the fully Bayes construction. As an aside, we show that SS-DL does not overfit in the sense that the posterior concentrates on smaller networks with fewer (up to the optimal number of) nodes and links. Our results provide new theoretical justifications for deep ReLU networks from a Bayesian point of view.

Analytic solution and stationary phase approximation for the Bayesian lasso and elastic net

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #49

Tom Michoel

The lasso and elastic net linear regression models impose a double-exponential prior distribution on the model parameters to achieve regression shrinkage and variable selection, allowing the inference of robust models from large data sets. However, there has been limited success in deriving estimates for the full posterior distribution of regression coefficients in these models, due to a need to evaluate analytically intractable partition function integrals. Here, the Fourier transform is used to express these integrals as complex-valued oscillatory integrals over "regression frequencies". This results in an analytic expansion and stationary phase approximation for the partition functions of the Bayesian lasso and elastic net, where the non-differentiability of the double-exponential prior has so far eluded such an approach. Use of this approximation leads to highly accurate numerical estimates for the expectation values and marginal posterior distributions of the regression coefficients, and allows for Bayesian inference of much higher dimensional models than previously possible.

Bayesian Model Selection Approach to Boundary Detection with Non-Local Priors

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #50

Fei Jiang · Guosheng Yin · Francesca Dominici

Based on non-local prior distributions, we propose a Bayesian model selection (BMS) procedure for boundary detection in a sequence of data with multiple systematic mean changes. The BMS method can effectively suppress the non-boundary spike points with large instantaneous changes. We speed up the algorithm by reducing the multiple change points to a series of single change point detection problems. We establish the consistency of the estimated number and locations of the change points under various prior distributions. Extensive simulation studies are conducted to compare the BMS with existing methods, and our approach is illustrated with application to the magnetic resonance imaging guided radiation therapy data.

Graphical model inference: Sequential Monte Carlo meets deterministic approximations

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #51

Fredrik Lindsten · Jouni Helske · Matti Vihola

Approximate inference in probabilistic graphical models (PGMs) can be grouped into deterministic methods and Monte-Carlo-based methods. The former can often provide accurate and rapid inferences, but are typically associated with biases that are hard to quantify. The latter enjoy asymptotic consistency, but can suffer from high computational costs. In this paper we present a way of bridging the gap between deterministic and stochastic inference. Specifically, we suggest an efficient sequential Monte Carlo (SMC) algorithm for PGMs which can leverage the output from deterministic inference methods. While generally applicable, we show explicitly how this can be done with loopy belief propagation, expectation propagation, and Laplace approximations. The resulting algorithm can be viewed as a post-correction of the biases associated with these methods and, indeed, numerical results show clear improvements over the baseline deterministic methods as well as over "plain" SMC.

Implicit Probabilistic Integrators for ODEs

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #52

Onur Teymur · Han Cheng Lie · Tim Sullivan · Ben Calderhead

We introduce a family of implicit probabilistic integrators for initial value problems (IVPs), taking as a starting point the multistep Adams–Moulton method. The implicit construction allows for dynamic feedback from the forthcoming time-step, in contrast to previous probabilistic integrators, all of which are based on explicit methods. We begin with a concise survey of the rapidly-expanding field of probabilistic ODE solvers. We then introduce our method, which builds on and adapts the work of Conrad et al. (2016) and Teymur et al. (2016), and provide a rigorous proof of its well-definedness and convergence. We discuss the problem of the calibration of such integrators and suggest one approach. We give an illustrative example highlighting the effect of the use of probabilistic integrators—including our new method—in the setting of parameter inference within an inverse problem.

A Bayes-Sard Cubature Method

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #53

Toni Karvonen · Chris J Oates · Simo Sarkka

This paper focusses on the formulation of numerical integration as an inferential task. To date, research effort has largely focussed on the development of Bayesian cubature, whose distributional output provides uncertainty quantification for the integral. However, the point estimators associated to Bayesian cubature can be inaccurate and acutely sensitive to the prior when the domain is high-dimensional. To address these drawbacks we introduce Bayes-Sard cubature, a probabilistic framework that combines the flexibility of Bayesian cubature with the robustness of classical cubatures which are well-established. This is achieved by considering a Gaussian process model for the integrand whose mean is a parametric regression model, with an improper prior on each regression coefficient. The features in the regression model consist of test functions which are guaranteed to be exactly integrated, with remaining degrees of freedom afforded to the non-parametric part. The asymptotic convergence of the Bayes-Sard cubature method is established and the theoretical results are numerically verified. In particular, we report two orders of magnitude reduction in error compared to Bayesian cubature in the context of a high-dimensional financial integral.

Deep State Space Models for Time Series Forecasting

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #54

Syama Sundar Rangapuram · Matthias W Seeger · Jan Gasthaus · Lorenzo Stella · Yuyang Wang · Tim Januschowski

We present a novel approach to probabilistic time series forecasting that combines state space models with deep learning. By parametrizing a per-time-series linear state space model with a jointly-learned recurrent neural network, our method retains desired properties of state space models such as data efficiency and interpretability, while making use of the ability to learn complex patterns from raw data offered by deep learning approaches. Our method scales gracefully from regimes where little training data is available to regimes where data from millions of time series can be leveraged to learn accurate models. We provide qualitative as well as quantitative results with the proposed method, showing that it compares favorably to the state-of-the-art.

BRUNO: A Deep Recurrent Model for Exchangeable Data

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #55

Iryna Korshunova · Jonas Degrave · Ferenc Huszar · Yarin Gal · Arthur Gretton · Joni Dambre

We present a novel model architecture which leverages deep learning tools to perform exact Bayesian inference on sets of high dimensional, complex observations. Our model is provably exchangeable, meaning that the joint distribution over observations is invariant under permutation: this property lies at the heart of Bayesian inference. The model does not require variational approximations to train, and new samples can be generated conditional on previous samples, with cost linear in the size of the conditioning set. The advantages of our architecture are demonstrated on learning tasks that require generalisation from short observed sequences while modelling sequence variability, such as conditional image generation, few-shot learning, and anomaly detection.

Scaling Gaussian Process Regression with Derivatives

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #56

David Eriksson · Kun Dong · Eric Lee · David Bindel · Andrew Wilson

Gaussian processes (GPs) with derivatives are useful in many applications, including Bayesian optimization, implicit surface reconstruction, and terrain reconstruction. Fitting a GP to function values and derivatives at \$n\$ points in \$d\$ dimensions requires linear solves and log determinants with an \${n(d+1) \times n(d+1)}\$ positive definite matrix-- leading to prohibitive \$\mathcal{O}(n^3d^3)\$ computations for standard direct methods. We propose iterative solvers using fast \$\mathcal{O}(nd)\$ matrix-vector multiplications (MVMs), together with pivoted Cholesky preconditioning that cuts the iterations to convergence by several orders of magnitude, allowing for fast kernel learning and prediction. Our approaches, together with dimensionality reduction, allows us to scale Bayesian optimization with derivatives to high-dimensional problems and large evaluation budgets.

Algebraic tests of general Gaussian latent tree models

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #57

Dennis Leung · Mathias Drton

We consider general Gaussian latent tree models in which the observed variables are not restricted to be leaves of the tree. Extending related recent work, we give a full semi-algebraic description of the set of covariance matrices of any such model. In other words, we find polynomial constraints that characterize when a matrix is the covariance matrix of a distribution in a given latent tree model. However, leveraging these constraints to test a given such model is often complicated by the number of constraints being large and by singularities of individual polynomials, which may invalidate standard approximations to relevant probability distributions. Illustrating with the star tree, we propose a new testing methodology that circumvents singularity issues by trading off some statistical estimation efficiency and handles cases with many constraints through recent advances on Gaussian approximation for maxima of sums of high-dimensional random vectors. Our test avoids the need to maximize the possibly multimodal likelihood function of such models and is applicable to models with larger number of variables. These points are illustrated in numerical experiments.

Differentially Private Bayesian Inference for Exponential Families

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #58

Garrett Bernstein · Daniel Sheldon

The study of private inference has been sparked by growing concern regarding the analysis of data when it stems from sensitive sources. We present the first method for private Bayesian inference in exponential families that properly accounts for noise introduced by the privacy mechanism. It is efficient because it works only with sufficient statistics and not individual data. Unlike other methods, it gives properly calibrated posterior beliefs in the non-asymptotic data regime.

Semi-crowdsourced Clustering with Deep Generative Models

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #59

Yucen Luo · TIAN TIAN · Jiaxin Shi · Jun Zhu · Bo Zhang

We consider the semi-supervised clustering problem where crowdsourcing provides noisy information about the pairwise comparisons on a small subset of data, i.e., whether a sample pair is in the same cluster. We propose a new approach that includes a deep generative model (DGM) to characterize low-level features of the data, and a statistical relational model for noisy pairwise annotations on its subset. The two parts share the latent variables. To make the model automatically trade-off between its complexity and fitting data, we also develop its fully Bayesian variant. The challenge of inference is addressed by fast (natural-gradient) stochastic variational inference algorithms, where we effectively combine variational message passing for the relational part and amortized learning of the DGM under a unified framework. Empirical results on synthetic and real-world datasets show that our model outperforms previous crowdsourced clustering methods.

Deep Poisson gamma dynamical systems

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #60

Dandan Guo · Bo Chen · Hao Zhang · Mingyuan Zhou

We develop deep Poisson-gamma dynamical systems (DPGDS) to model sequentially observed multivariate count data, improving previously proposed models by not only mining deep hierarchical latent structure from the data, but also capturing both first-order and long-range temporal dependencies. Using sophisticated but simple-to-implement data augmentation techniques, we derived closed-form Gibbs sampling update equations by first backward and upward propagating auxiliary latent counts, and then forward and downward sampling latent variables. Moreover, we develop stochastic gradient MCMC inference that is scalable to very long multivariate count time series. Experiments on both synthetic and a variety of real-world data demonstrate that the proposed model not only has excellent predictive performance, but also provides highly interpretable multilayer latent structure to represent hierarchical and temporal information propagation.

Deep State Space Models for Unconditional Word Generation

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #61

Florian Schmidt · Thomas Hofmann

Autoregressive feedback is considered a necessity for successful unconditional text generation using stochastic sequence models. However, such feedback is known to introduce systematic biases into the training process and it obscures a principle of generation: committing to global information and forgetting local nuances. We show that a non-autoregressive deep state space model with a clear separation of global and local uncertainty can be built from only two ingredients: An independent noise source and a deterministic transition function. Recent advances on flow-based variational inference can be used to train an evidence lower-bound without resorting to annealing, auxiliary losses or similar measures. The result is a highly interpretable generative model on par with comparable auto-regressive models on the task of word generation.

Modular Networks: Learning to Decompose Neural Computation

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #62

Louis Kirsch · Julius Kunze · David Barber

Scaling model capacity has been vital in the success of deep learning. For a typical network, necessary compute resources and training time grow dramatically with model size. Conditional computation is a promising way to increase the number of parameters with a relatively small increase in resources. We propose a training algorithm that flexibly chooses neural modules based on the data to be processed. Both the decomposition and modules are learned end-to-end. In contrast to existing approaches, training does not rely on regularization to enforce diversity in module use. We apply modular networks both to image recognition and language modeling tasks, where we achieve superior performance compared to several baselines. Introspection reveals that modules specialize in interpretable contexts.

Gaussian Process Prior Variational Autoencoders

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #63

Francesco Paolo Casale · Adrian Dalca · Luca Saglietti · Jennifer Listgarten · Nicolo Fusi

Variational autoencoders (VAE) are a powerful and widely-used class of models to learn complex data distributions in an unsupervised fashion. One important limitation of VAEs is the prior assumption that latent sample representations are independent and identically distributed. However, for many important datasets, such as time-series of images, this assumption is too strong: accounting for covariances between samples, such as those in time, can yield to a more appropriate model specification and improve performance in downstream tasks. In this work, we introduce a new model, the Gaussian Process (GP) Prior Variational Autoencoder (GPPVAE), to specifically address this issue. The GPPVAE aims to combine the power of VAEs with the ability to model correlations afforded by GP priors. To achieve efficient inference in this new class of models, we leverage structure in the covariance matrix, and introduce a new stochastic backpropagation strategy that allows for computing stochastic gradients in a distributed and low-memory fashion. We show that our method outperforms conditional VAEs (CVAEs) and an adaptation of standard VAEs in two image data applications.

Bayesian Semi-supervised Learning with Graph Gaussian Processes

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #64

Yin Cheng Ng · Nicolò Colombo · Ricardo Silva

We propose a data-efficient Gaussian process-based Bayesian approach to the semi-supervised learning problem on graphs. The proposed model shows extremely competitive performance when compared to the state-of-the-art graph neural networks on semi-supervised learning benchmark experiments, and outperforms the neural networks in active learning experiments where labels are scarce. Furthermore, the model does not require a validation data set for early stopping to control over-fitting. Our model can be viewed as an instance of empirical distribution regression weighted locally by network connectivity. We further motivate the intuitive construction of the model with a Bayesian linear model interpretation where the node features are filtered by an operator related to the graph Laplacian. The method can be easily implemented by adapting off-the-shelf scalable variational inference algorithms for Gaussian processes.

Inference in Deep Gaussian Processes using Stochastic Gradient Hamiltonian Monte Carlo

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #65

Marton Havasi · José Miguel Hernández-Lobato · Juan José Murillo-Fuentes

Deep Gaussian Processes (DGPs) are hierarchical generalizations of Gaussian Processes that combine well calibrated uncertainty estimates with the high flexibility of multilayer models. One of the biggest challenges with these models is that exact inference is intractable. The current state-of-the-art inference method, Variational Inference (VI), employs a Gaussian approximation to the posterior distribution. This can be a potentially poor unimodal approximation of the generally multimodal posterior. In this work, we provide evidence for the non-Gaussian nature of the posterior and we apply the Stochastic Gradient Hamiltonian Monte Carlo method to generate samples. To efficiently optimize the hyperparameters, we introduce the Moving Window MCEM algorithm. This results in significantly better predictions at a lower computational cost than its VI counterpart. Thus our method establishes a new state-of-the-art for inference in DGPs.

Variational Bayesian Monte Carlo

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #66

Luigi Acerbi

Many probabilistic models of interest in scientific computing and machine learning have expensive, black-box likelihoods that prevent the application of standard techniques for Bayesian inference, such as MCMC, which would require access to the gradient or a large number of likelihood evaluations. We introduce here a novel sample-efficient inference framework, Variational Bayesian Monte Carlo (VBMC). VBMC combines variational inference with Gaussian-process based, active-sampling Bayesian quadrature, using the latter to efficiently approximate the intractable integral in the variational objective. Our method produces both a nonparametric approximation of the posterior distribution and an approximate lower bound of the model evidence, useful for model selection. We demonstrate VBMC both on several synthetic likelihoods and on a neuronal model with data from real neurons. Across all tested problems and dimensions (up to D = 10), VBMC performs consistently well in reconstructing the posterior and the model evidence with a limited budget of likelihood evaluations, unlike other methods that work only in very low dimensions. Our framework shows great promise as a novel tool for posterior and model inference with expensive, black-box likelihoods.

Bayesian Alignments of Warped Multi-Output Gaussian Processes

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #67

Markus Kaiser · Clemens Otte · Thomas Runkler · Carl Henrik Ek

We propose a novel Bayesian approach to modelling nonlinear alignments of time series based on latent shared information. We apply the method to the real-world problem of finding common structure in the sensor data of wind turbines introduced by the underlying latent and turbulent wind field. The proposed model allows for both arbitrary alignments of the inputs and non-parametric output warpings to transform the observations. This gives rise to multiple deep Gaussian process models connected via latent generating processes. We present an efficient variational approximation based on nested variational compression and show how the model can be used to extract shared information between dependent time series, recovering an interpretable functional decomposition of the learning problem. We show results for an artificial data set and real-world data of two wind turbines.

Automating Bayesian optimization with Bayesian optimization

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #68

Gustavo Malkomes · Roman Garnett

Bayesian optimization is a powerful tool for global optimization of expensive functions. One of its key components is the underlying probabilistic model used for the objective function f. In practice, however, it is often unclear how one should appropriately choose a model, especially when gathering data is expensive. In this work, we introduce a novel automated Bayesian optimization approach that dynamically selects promising models for explaining the observed data using Bayesian Optimization in the model space. Crucially, we account for the uncertainty in the choice of model; our method is capable of using multiple models to represent its current belief about f and subsequently using this information for decision making. We argue, and demonstrate empirically, that our approach automatically finds suitable models for the objective function, which ultimately results in more-efficient optimization.

Infinite-Horizon Gaussian Processes

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #69

Arno Solin · James Hensman · Richard E Turner

Gaussian processes provide a flexible framework for forecasting, removing noise, and interpreting long temporal datasets. State space modelling (Kalman filtering) enables these non-parametric models to be deployed on long datasets by reducing the complexity to linear in the number of data points. The complexity is still cubic in the state dimension m which is an impediment to practical application. In certain special cases (Gaussian likelihood, regular spacing) the GP posterior will reach a steady posterior state when the data are very long. We leverage this and formulate an inference scheme for GPs with general likelihoods, where inference is based on single-sweep EP (assumed density filtering). The infinite-horizon model tackles the cubic cost in the state dimensionality and reduces the cost in the state dimension m to O(m^2) per data point. The model is extended to online-learning of hyperparameters. We show examples for large finite-length modelling problems, and present how the method runs in real-time on a smartphone on a continuous data stream updated at 100 Hz.

Learning Gaussian Processes by Minimizing PAC-Bayesian

Generalization Bounds

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #70

David Reeb · Andreas Doerr · Sebastian Gerwinn · Barbara Rakitsch

Gaussian Processes (GPs) are a generic modelling tool for supervised learning. While they have been successfully applied on large datasets, their use in safety-critical applications is hindered by the lack of good performance guarantees. To this end, we propose a method to learn GPs and their sparse approximations by directly optimizing a PAC-Bayesian bound on their generalization performance, instead of maximizing the marginal likelihood. Besides its theoretical appeal, we find in our evaluation that our learning method is robust and yields significantly better generalization guarantees than other common GP approaches on several regression benchmark datasets.

Algorithmic Linearly Constrained Gaussian Processes

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #71

Markus Lange-Hegermann

We algorithmically construct multi-output Gaussian process priors which satisfy linear differential equations. Our approach attempts to parametrize all solutions of the equations using Gröbner bases. If successful, a push forward Gaussian process along the paramerization is the desired prior. We consider several examples from physics, geomathmatics and control, among them the full inhomogeneous system of Maxwell's equations. By bringing together stochastic learning and computeralgebra in a novel way, we combine noisy observations with precise algebraic computations.

Efficient Projection onto the Perfect Phylogeny Model

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #72

Bei Jia · Surjyendu Ray · Sam Safavi · José Bento

Several algorithms build on the perfect phylogeny model to infer evolutionary trees. This problem is particularly hard when evolutionary trees are inferred from the fraction of genomes that have mutations in different positions, across different samples. Existing algorithms might do extensive searches over the space of possible trees. At the center of these algorithms is a projection problem that assigns a fitness cost to phylogenetic trees. In order to perform a wide search over the space of the trees, it is critical to solve this projection problem fast. In this paper, we use Moreau's decomposition for proximal operators, and a tree reduction scheme, to develop a new algorithm to compute this projection. Our algorithm terminates with an exact solution in a finite number of steps, and is extremely fast. In particular, it can search over all evolutionary trees with fewer than 11 nodes, a size relevant for several biological problems (more than 2 billion trees) in about 2 hours.

Distributed k-Clustering for Data with Heavy Noise

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #73

Shi Li · Xiangyu Guo

In this paper, we consider the k-center/median/means clustering with outliers problems (or the (k, z)-center/median/means problems) in the distributed setting. Most previous distributed algorithms have their communication costs linearly depending on z, the number of outliers. Recently Guha et al.[10] overcame this dependence issue by considering bi-criteria approximation algorithms that output solutions with $2z$ outliers. For the case where z is large, the extra z outliers discarded by the algorithms might be too large, considering that the data gathering process might be costly. In this paper, we improve the number of outliers to the best possible $(1+\epsilon)z$, while maintaining the $O(1)$-approximation ratio and independence of communication cost on z. The problems we consider include the (k, z)-center problem, and (k, z)-median/means problems in Euclidean metrics. Implementation of the our algorithm for (k, z)-center shows that it outperforms many previous algorithms, both in terms of the communication cost and quality of the output solution.

Communication Compression for Decentralized Training

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #74

Hanlin Tang · Shaoduo Gan · Ce Zhang · Tong Zhang · Ji Liu

Optimizing distributed learning systems is an art of balancing between computation and communication. There have been two lines of research that try to deal with slower networks: {\em communication compression} for low bandwidth networks, and {\em decentralization} for high latency networks. In this paper, We explore a natural question: {\em can the combination of both techniques lead to a system that is robust to both bandwidth and latency?}

Although the system implication of such combination is trivial, the underlying theoretical principle and algorithm design is challenging: unlike centralized algorithms, simply compressing {\rc exchanged information, even in an unbiased stochastic way, within the decentralized network would accumulate the error and cause divergence.} In this paper, we develop a framework of quantized, decentralized training and propose two different strategies, which we call {\em extrapolation compression} and {\em difference compression}. We analyze both algorithms and prove both converge at the rate of $O(1/\sqrt{nT})$ where n is the number of workers and T is the number of iterations, matching the convergence rate for full precision, centralized training. We validate our algorithms and find that our proposed algorithm outperforms the best of merely decentralized and merely quantized algorithm significantly for networks with {\em both} high latency and low bandwidth.

Do Less, Get More: Streaming Submodular Maximization with Subsampling

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #75

Moran Feldman · Amin Karbasi · Ehsan Kazemi

In this paper, we develop the first one-pass streaming algorithm for submodular maximization that does not evaluate the entire stream even once. By carefully subsampling each element of the data stream, our algorithm enjoys the tightest approximation guarantees in various settings while having the smallest memory footprint and requiring the lowest number of function evaluations. More specifically, for a monotone submodular function and a p-matchoid constraint, our randomized algorithm achieves a $4p$ approximation ratio (in expectation) with $O(k)$ memory and $O(km/p)$ queries per element (k is the size of the largest feasible solution and m is the number of matroids used to define the constraint). For the non-monotone case, our approximation ratio increases only slightly to $4p+2-o(1)$. To the best or our knowledge, our algorithm is the first that combines the benefits of streaming and subsampling in a novel way in order to truly scale submodular maximization to massive machine learning problems. To showcase its practicality, we empirically evaluated the performance of our algorithm on a video summarization application and observed that it outperforms the state-of-the-art algorithm by up to fifty-fold while maintaining practically the same utility. We also evaluated the scalability of our algorithm on a large dataset of Uber pick up locations.

Optimal Algorithms for Continuous Non-monotone Submodular and DR-Submodular Maximization

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #76

Rad Niazadeh · Tim Roughgarden · Joshua Wang

In this paper we study the fundamental problems of maximizing a continuous non monotone submodular function over a hypercube, with and without coordinate-wise concavity. This family of optimization problems has several applications in machine learning, economics, and communication systems. Our main result is the first 1/2 approximation algorithm for continuous submodular function maximization; this approximation factor of is the best possible for algorithms that use only polynomially many queries. For the special case of DR-submodular maximization, we provide a faster

1/2-approximation algorithm that runs in (almost) linear time. Both of these results improve upon prior work [Bian et al., 2017, Soma and Yoshida, 2017, Buchbinder et al., 2012]. Our first algorithm is a single-pass algorithm that uses novel ideas such as reducing the guaranteed approximation problem to analyzing a zero-sum game for each coordinate, and incorporates the geometry of this zero-sum game to fix the value at this coordinate. Our second algorithm is a faster single-pass algorithm that exploits coordinate-wise concavity to identify a monotone equilibrium condition sufficient for getting the required approximation guarantee, and hunts for the equilibrium point using binary search. We further run experiments to verify the performance of our proposed algorithms in related machine learning applications.

Provable Variational Inference for Constrained Log-Submodular Models

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #77

Josip Djolonga · Stefanie Jegelka · Andreas Krause

Submodular maximization problems appear in several areas of machine learning and data science, as many useful modelling concepts such as diversity and coverage satisfy this natural diminishing returns property. Because the data defining these functions, as well as the decisions made with the computed solutions, are subject to statistical noise and randomness, it is arguably necessary to go beyond computing a single approximate optimum and quantify its inherent uncertainty. To this end, we define a rich class of probabilistic models associated with constrained submodular maximization problems. These capture log-submodular dependencies of arbitrary order between the variables, but also satisfy hard combinatorial constraints. Namely, the variables are assumed to take on one of — possibly exponentially many — set of states, which form the bases of a matroid. To perform inference in these models we design novel variational inference algorithms, which carefully leverage the combinatorial and probabilistic properties of these objects. In addition to providing completely tractable and well-understood variational approximations, our approach results in the minimization of a convex upper bound on the log-partition function. The bound can be efficiently evaluated using greedy algorithms and optimized using any first-order method. Moreover, for the case of facility location and weighted coverage functions, we prove the first constant factor guarantee in this setting — an efficiently certifiable e/(e-1) approximation of the log-partition function. Finally, we empirically demonstrate the effectiveness of our approach on several instances.

Fast greedy algorithms for dictionary selection with generalized sparsity constraints

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #78

Kaito Fujii · Tasuku Soma

In dictionary selection, several atoms are selected from finite candidates that successfully approximate given data points in the sparse representation. We propose a novel efficient greedy algorithm for dictionary selection. Not only does our algorithm work much faster than the known methods, but it can also handle more complex sparsity constraints, such as average sparsity. Using numerical experiments, we show that our algorithm outperforms the known methods for dictionary selection, achieving competitive performances with dictionary learning algorithms in a smaller running time.

Boolean Decision Rules via Column Generation

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #79

Sanjeeb Dash · Oktay Gunluk · Dennis Wei

This paper considers the learning of Boolean rules in either disjunctive normal form (DNF, OR-of-ANDs, equivalent to decision rule sets) or conjunctive normal form (CNF, AND-of-ORs) as an interpretable model for classification. An integer program is formulated to optimally trade classification accuracy for rule simplicity. Column generation (CG) is used to efficiently search over an exponential number of candidate clauses (conjunctions or disjunctions) without the need for heuristic rule mining. This approach also bounds the gap between the selected rule set and the best possible rule set on the training data. To handle large datasets, we propose an approximate CG algorithm using randomization. Compared to three recently proposed alternatives, the CG algorithm dominates the accuracy-simplicity trade-off in 8 out of 16 datasets. When maximized for accuracy, CG is competitive with rule learners designed for this purpose, sometimes finding significantly simpler solutions that are no less accurate.

Computing Kantorovich-Wasserstein Distances on d-dimensional histograms using $(d+1)$-partite graphs

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #80

Gennaro Auricchio · Federico Bassetti · Stefano Gualandi · Marco Veneroni

This paper presents a novel method to compute the exact Kantorovich-Wasserstein distance between a pair of d-dimensional histograms having n bins each. We prove that this problem is equivalent to an uncapacitated minimum cost flow problem on a $(d+1)$-partite graph with $(d+1)n$ nodes and $dn^{\frac{d+1}{d}}$ arcs, whenever the cost is separable along the principal d-dimensional directions. We show numerically the benefits of our approach by computing the Kantorovich-Wasserstein distance of order 2 among two sets of instances: gray scale images and

d-dimensional biomedical histograms. On these types of instances, our approach is competitive with state-of-the-art optimal transport algorithms.

Adaptive Negative Curvature Descent with Applications in Non-convex Optimization

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #81

Mingrui Liu · Zhe Li · Xiaoyu Wang · Jinfeng Yi · Tianbao Yang

Negative curvature descent (NCD) method has been utilized to design deterministic or stochastic algorithms for non-convex optimization aiming at finding second-order stationary points or local minima. In existing studies, NCD needs to approximate the smallest eigen-value of the Hessian matrix with a sufficient precision (e.g., $\epsilon_2\ll 1$) in order to achieve a sufficiently accurate second-order stationary solution (i.e., $\lambda_{\min}(\nabla^2 f(\x))\geq -\epsilon_2)$. One issue with this approach is that the target precision ϵ_2 is usually set to be very small in order to find a high quality solution, which increases the complexity for computing a negative curvature. To address this issue, we propose an adaptive NCD to allow for an adaptive error dependent on the current gradient's magnitude in approximating the smallest eigen-value of the Hessian, and to encourage competition between a noisy NCD step and gradient descent step. We consider the applications of the proposed adaptive NCD for both deterministic and stochastic non-convex optimization, and demonstrate that it can help reduce the the overall complexity in computing the negative curvatures during the course of optimization without sacrificing the iteration complexity.

Implicit Bias of Gradient Descent on Linear Convolutional Networks

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #82

Suriya Gunasekar · Jason Lee · Daniel Soudry · Nati Srebro

We show that gradient descent on full-width linear convolutional networks of depth L converges to a linear predictor related to the $\ell_{2/L}$ bridge penalty in the frequency domain. This is in contrast to linearly fully connected networks, where gradient descent converges to the hard margin linear SVM solution, regardless of depth.

Deep Generative Models for Distribution-Preserving Lossy

Compression

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #83

Michael Tschannen · Eirikur Agustsson · Mario Lucic

We propose and study the problem of distribution-preserving lossy compression. Motivated by recent advances in extreme image compression which allow to maintain artifact-free reconstructions even at very low bitrates, we propose to optimize the rate-distortion tradeoff under the constraint that the reconstructed samples follow the distribution of the training data. The resulting compression system recovers both ends of the spectrum: On one hand, at zero bitrate it learns a generative model of the data, and at high enough bitrates it achieves perfect reconstruction. Furthermore, for intermediate bitrates it smoothly interpolates between learning a generative model of the training data and perfectly reconstructing the training samples. We study several methods to approximately solve the proposed optimization problem, including a novel combination of Wasserstein GAN and Wasserstein Autoencoder, and present an extensive theoretical and empirical characterization of the proposed compression systems.

Visual Object Networks: Image Generation with Disentangled 3D Representations

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #84

Jun-Yan Zhu · Zhoutong Zhang · Chengkai Zhang · Jiajun Wu · Antonio Torralba · Josh Tenenbaum · Bill Freeman

Recent progress in deep generative models has led to tremendous breakthroughs in image generation. While being able to synthesize photorealistic images, existing models lack an understanding of our underlying 3D world. Different from previous works built on 2D datasets and models, we present a new generative model, Visual Object Networks (VON), synthesizing natural images of objects with a disentangled 3D representation. Inspired by classic graphics rendering pipelines, we unravel the image formation process into three conditionally independent factors---viewpoint, shape, and texture---and present an end-to-end adversarial learning framework that jointly models 3D shape and 2D texture. Our model first learns to synthesize 3D shapes that are indistinguishable from real shapes. It then renders the object's 2.5D sketches (i.e., silhouette and depth map) from its shape under a sampled viewpoint. Finally, it learns to add realistic textures to these 2.5D sketches to generate realistic images. The VON not only generates images that are more realistic than the state-of-the-art 2D image synthesis methods but also enables many 3D operations such as changing the viewpoint of a generated image, shape and texture editing, and linear interpolation in texture and shape space.

Nonlocal Neural Networks, Nonlocal Diffusion and Nonlocal Modeling

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #85

Yunzhe Tao · Qi Sun · Qiang Du · Wei Liu

Nonlocal neural networks have been proposed and shown to be effective in several computer vision tasks, where the nonlocal operations can directly capture long-range dependencies in the feature space. In this paper, we study the nature of diffusion and damping effect of nonlocal networks by doing spectrum analysis on the weight matrices of the well-trained networks, and then propose a new formulation of the nonlocal block. The new block not only learns the nonlocal interactions but also has stable dynamics, thus allowing deeper nonlocal structures. Moreover, we interpret our formulation from the general nonlocal modeling perspective, where we make connections between the proposed nonlocal network and other nonlocal models, such as nonlocal diffusion process and Markov jump process.

Can We Gain More from Orthogonality Regularizations in Training Deep Networks?

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #86

Nitin Bansal · Xiaohan Chen · Zhangyang Wang

This paper seeks to answer the question: as the (near-) orthogonality of weights is found to be a favorable property for training deep convolutional neural networks, how can we enforce it in more effective and easy-to-use ways? We develop novel orthogonality regularizations on training deep CNNs, utilizing various advanced analytical tools such as mutual coherence and restricted isometry property. These plug-and-play regularizations can be conveniently incorporated into training almost any CNN without extra hassle. We then benchmark their effects on state-of-the-art models: ResNet, WideResNet, and ResNeXt, on several most popular computer vision datasets: CIFAR-10, CIFAR-100, SVHN and ImageNet. We observe consistent performance gains after applying those proposed regularizations, in terms of both the final accuracies achieved, and faster and more stable convergences. We have made our codes and pre-trained models publicly available: https://github.com/nbansal90/Can-we-Gain-More-from-Orthogonality.

Discrimination-aware Channel Pruning for Deep Neural Networks

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #87

Zhuangwei Zhuang · Mingkui Tan · Bohan Zhuang · Jing Liu · Yong Guo · Qingyao Wu · Junzhou Huang · Jinhui Zhu

Channel pruning is one of the predominant approaches for deep model compression. Existing pruning methods either train from scratch with sparsity constraints on channels, or minimize the reconstruction error between the pre-trained feature maps and the compressed ones. Both strategies suffer from some limitations: the former kind is computationally expensive and difficult to converge, whilst the latter kind optimizes the reconstruction error but ignores the discriminative power of channels. To overcome these drawbacks, we investigate a simple-yet-effective method, called discrimination-aware channel pruning, to choose those channels that really contribute to discriminative power. To this end, we introduce additional losses into the network to increase the discriminative power of intermediate layers and then select the most discriminative channels for each layer by considering the additional loss and the reconstruction error. Last, we propose a greedy algorithm to conduct channel selection and parameter optimization in an iterative way. Extensive experiments demonstrate the effectiveness of our method. For example, on ILSVRC-12, our pruned ResNet-50 with 30% reduction of channels even outperforms the original model by 0.39% in top-1 accuracy.

Probabilistic Model-Agnostic Meta-Learning

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #88

Chelsea Finn · Kelvin Xu · Sergey Levine

Meta-learning for few-shot learning entails acquiring a prior over previous tasks and experiences, such that new tasks be learned from small amounts of data. However, a critical challenge in few-shot learning is task ambiguity: even when a powerful prior can be meta-learned from a large number of prior tasks, a small dataset for a new task can simply be too ambiguous to acquire a single model (e.g., a classifier) for that task that is accurate. In this paper, we propose a probabilistic meta-learning algorithm that can sample models for a new task from a model distribution. Our approach extends model-agnostic meta-learning, which adapts to new tasks via gradient descent, to incorporate a parameter distribution that is trained via a variational lower bound. At meta-test time, our algorithm adapts via a simple procedure that injects noise into gradient descent, and at meta-training time, the model is trained such that this stochastic adaptation procedure produces samples from the approximate model posterior. Our experimental results show that our method can sample plausible classifiers and regressors in ambiguous few-shot learning problems. We also show how reasoning about ambiguity can also be used for downstream active learning problems.

FastGRNN: A Fast, Accurate, Stable and Tiny Kilobyte Sized

Gated Recurrent Neural Network

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #89

Aditya Kusupati · Manish Singh · Kush Bhatia · Ashish Kumar · Prateek Jain · Manik Varma

This paper develops the FastRNN and FastGRNN algorithms to address the twin RNN limitations of inaccurate training and inefficient prediction. Previous approaches have improved accuracy at the expense of prediction costs making them infeasible for resource-constrained and real-time applications. Unitary RNNs have increased accuracy somewhat by restricting the range of the state transition matrix's singular values but have also increased the model size as they require a larger number of hidden units to make up for the loss in expressive power. Gated RNNs have obtained state-of-the-art accuracies by adding extra parameters thereby resulting in even larger models. FastRNN addresses these limitations by adding a residual connection that does not constrain the range of the singular values explicitly and has only two extra scalar parameters. FastGRNN then extends the residual connection to a gate by reusing the RNN matrices to match state-of-the-art gated RNN accuracies but with a 2-4x smaller model. Enforcing FastGRNN's matrices to be low-rank, sparse and quantized resulted in accurate models that could be up to 35x smaller than leading gated and unitary RNNs. This allowed FastGRNN to accurately recognize the "Hey Cortana" wakeword with a 1 KB model and to be deployed on severely resource-constrained IoT microcontrollers too tiny to store other RNN models. FastGRNN's code is available at (https://github.com/Microsoft/EdgeML/).

Understanding Batch Normalization

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #90

Nils Bjorck · Carla P Gomes · Bart Selman · Kilian Weinberger

Batch normalization (BN) is a technique to normalize activations in intermediate layers of deep neural networks. Its tendency to improve accuracy and speed up training have established BN as a favorite technique in deep learning. Yet, despite its enormous success, there remains little consensus on the exact reason and mechanism behind these improvements. In this paper we take a step towards a better understanding of BN, following an empirical approach. We conduct several experiments, and show that BN primarily enables training with larger learning rates, which is the cause for faster convergence and better generalization. For networks without BN we demonstrate how large gradient updates can result in diverging loss and activations growing uncontrollably with network depth, which limits possible learning rates. BN avoids this problem by constantly correcting activations to be zero-mean and of unit standard deviation, which enables larger gradient steps, yields faster convergence and may help bypass sharp local minima. We further show various ways in which gradients and activations of deep unnormalized networks are ill-behaved. We contrast our results against recent findings in random matrix theory, shedding new light on classical initialization schemes and their consequences.

How Many Samples are Needed to Estimate a Convolutional Neural Network?

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #91

Simon Du · Yining Wang · Xiyu Zhai · Sivaraman Balakrishnan · Ruslan Salakhutdinov · Aarti Singh

A widespread folklore for explaining the success of Convolutional Neural Networks (CNNs) is that CNNs use a more compact representation than the Fully-connected Neural Network (FNN) and thus require fewer training samples to accurately estimate their parameters. We initiate the study of rigorously characterizing the sample complexity of estimating CNNs. We show that for an m-dimensional convolutional filter with linear activation acting on a d-dimensional input, the sample complexity of achieving population prediction error of ϵ is $\widetilde{O(m/\epsilon^2)$, whereas the sample-complexity for its FNN counterpart is lower bounded by $\Omega(d/\epsilon^2)$ samples. Since, in typical settings $m \ll d$, this result demonstrates the advantage of using a CNN. We further consider the sample complexity of estimating a one-hidden-layer CNN with linear activation where both the m-dimensional convolutional filter and the r-dimensional output weights are unknown. For this model, we show that the sample complexity is $\widetilde{O}\left((m+r)/\epsilon^2\right)$ when the ratio between the stride size and the filter size is a constant. For both models, we also present lower bounds showing our sample complexities are tight up to logarithmic factors. Our main tools for deriving these results are a localized empirical process analysis and a new lemma characterizing the convolutional structure. We believe that these tools may inspire further developments in understanding CNNs.

Robust Detection of Adversarial Attacks by Modeling the Intrinsic Properties of Deep Neural Networks

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #92

Zhihao Zheng · Pengyu Hong

It has been shown that deep neural network (DNN) based classifiers are vulnerable to human-imperceptive adversarial perturbations which can cause DNN classifiers to output wrong predictions with high confidence. We propose an unsupervised learning approach to detect adversarial inputs without any knowledge of attackers. Our approach tries to capture the intrinsic properties of a DNN classifier and uses them to detect adversarial inputs. The intrinsic properties used in this study are the output distributions of the hidden neurons in a DNN classifier presented with natural images. Our approach can be easily applied to any DNN classifiers or combined with other defense strategy to improve robustness. Experimental results show that our approach demonstrates state-of-the-art robustness in defending black-box and gray-box attacks.

Combinatorial Optimization with Graph Convolutional Networks and Guided Tree Search

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #93

Zhuwen Li · Qifeng Chen · Vladlen Koltun

We present a learning-based approach to computing solutions for certain NP-hard problems. Our approach combines deep learning techniques with useful algorithmic elements from classic heuristics. The central component is a graph convolutional network that is trained to estimate the likelihood, for each vertex in a graph, of whether this vertex is part of the optimal solution. The network is designed and trained to synthesize a diverse set of solutions, which enables rapid exploration of the solution space via tree search. The presented approach is evaluated on four canonical NP-hard problems and five datasets, which include benchmark satisfiability problems and real social network graphs with up to a hundred thousand nodes. Experimental results demonstrate that the presented approach substantially outperforms recent deep learning work, and performs on par with highly optimized state-of-the-art heuristic solvers for some NP-hard problems. Experiments indicate that our approach generalizes across datasets, and scales to graphs that are orders of magnitude larger than those used during training.

Automatic differentiation in ML: Where we are and where we should be going

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #94

Bart van Merrienboer · Olivier Breuleux · Arnaud Bergeron · Pascal Lamblin

We review the current state of automatic differentiation (AD) for array programming in machine learning (ML), including the different approaches such as operator overloading (OO) and source transformation (ST) used for AD, graph-based intermediate representations for programs, and source languages. Based on these insights, we introduce a new graph-based intermediate representation (IR) which specifically aims to efficiently support fully-general AD for array programming. Unlike existing dataflow programming representations in ML frameworks, our IR naturally supports function calls, higher-order functions and recursion, making ML models easier to implement. The ability to represent closures allows us to perform AD using ST without a tape, making the resulting derivative (adjoint) program amenable to ahead-of-time optimization using tools from functional language compilers, and enabling higher-order derivatives. Lastly, we introduce a proof of concept compiler toolchain called Myia which uses a subset of Python as a front end.

Realistic Evaluation of Deep Semi-Supervised Learning Algorithms

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #95

Avital Oliver · Augustus Odena · Colin A Raffel · Ekin Dogus Cubuk · Ian Goodfellow

Semi-supervised learning (SSL) provides a powerful framework for leveraging unlabeled data when labels are limited or expensive to obtain. SSL algorithms based on deep neural networks have recently proven successful on standard benchmark tasks. However, we argue that these benchmarks fail to address many issues that SSL algorithms would face in real-world applications. After creating a unified reimplementation of various widely-used SSL techniques, we test them in a suite of experiments designed to address these issues. We find that the performance of simple baselines which do not use unlabeled data is often underreported, SSL methods differ in sensitivity to the amount of labeled and unlabeled data, and performance can degrade substantially when the unlabeled dataset contains out-of-distribution examples. To help guide SSL research towards real-world applicability, we make our unified reimplemention and evaluation platform publicly available.

Toddler-Inspired Visual Object Learning

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #96

Sven Bambach · David Crandall · Linda Smith · Chen Yu

Real-world learning systems have practical limitations on the quality and quantity of the training datasets that they can collect and consider. How should a system go about choosing a subset of the possible training examples that still allows for learning accurate, generalizable models? To help address this question, we draw inspiration from a highly efficient practical learning system: the human child. Using head-mounted cameras, eye gaze trackers, and a model of foveated vision, we collected first-person (egocentric) images that represents a highly accurate approximation of the "training data" that toddlers' visual systems collect in everyday, naturalistic learning contexts. We used state-of-the-art computer vision learning models (convolutional neural networks) to help characterize the structure of these data, and found that child data produce significantly better object models than egocentric data experienced by adults in exactly the same environment. By using the CNNs as a modeling tool to investigate the properties of the child data that may enable this rapid learning, we found that child data exhibit a unique combination of quality and diversity, with not only many similar large, high-quality object views but also a greater number and diversity of rare views. This novel methodology of analyzing the visual "training data" used by children may not only reveal insights to improve machine learning, but also may suggest new experimental tools to better understand infant learning in developmental psychology.

Generalisation in humans and deep neural networks

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #97

Robert Geirhos · Carlos R. M. Temme · Jonas Rauber · Heiko H. Schütt · Matthias Bethge · Felix A. Wichmann

We compare the robustness of humans and current convolutional deep neural networks (DNNs) on object recognition under twelve different types of image degradations. First, using three well known DNNs (ResNet-152, VGG-19, GoogLeNet) we find the human visual system to be more robust to nearly all of the tested image manipulations, and we observe progressively diverging classification error-patterns between humans and DNNs when the signal gets weaker. Secondly, we show that DNNs trained directly on distorted images consistently surpass human performance on the exact distortion types they were trained on, yet they display extremely poor generalisation abilities when tested on other distortion types. For example, training on salt-and-pepper noise does not imply robustness on uniform white noise and vice versa. Thus, changes in the noise distribution between training and testing constitutes a crucial challenge to deep learning vision systems that can be systematically addressed in a lifelong machine learning approach. Our new dataset consisting of 83K carefully measured human psychophysical trials provide a useful reference for lifelong robustness against image degradations set by the human visual system.

Assessing the Scalability of Biologically-Motivated Deep Learning Algorithms and Architectures

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #98

Sergey Bartunov · Adam Santoro · Blake Richards · Luke Marris · Geoffrey E Hinton · Timothy Lillicrap

The backpropagation of error algorithm (BP) is impossible to implement in a real brain. The recent success of deep networks in machine learning and AI, however, has inspired proposals for understanding how the brain might learn across multiple layers, and hence how it might approximate BP. As of yet, none of these proposals have been rigorously evaluated on tasks where BP-guided deep learning has proved critical, or in architectures more structured than simple fully-connected networks. Here we present results on scaling up biologically motivated models of deep learning on datasets which need deep networks with appropriate architectures to achieve good performance. We present results on the MNIST, CIFAR-10, and ImageNet datasets and explore variants of target-propagation (TP) and feedback alignment (FA) algorithms, and explore performance in both fully- and locally-connected architectures. We also introduce weight-transport-free variants of difference target propagation (DTP) modified to remove backpropagation from the penultimate layer. Many of these algorithms perform well for MNIST, but for CIFAR and ImageNet we find that TP and FA variants perform significantly worse than BP, especially for networks composed of locally connected units, opening questions about whether new architectures and algorithms are required to scale these approaches. Our results and implementation details help establish baselines for biologically motivated deep learning schemes going forward.

Incorporating Context into Language Encoding Models for fMRI

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #99

Shailee Jain · Alexander Huth

Language encoding models help explain language processing in the human brain by learning functions that predict brain responses from the language stimuli that elicited them. Current word embedding-based approaches treat each stimulus word independently and thus ignore the influence of context on language understanding. In this work we instead build encoding models using rich contextual representations derived from an LSTM language model. Our models show a significant improvement in encoding performance relative to state-of-the-art embeddings in nearly every brain area. By varying the amount of context used in the models and providing the models with distorted context, we show that this improvement is due to a combination of better word embeddings learned by the LSTM language model and contextual information. We are also able to use our models to map context sensitivity across the cortex. These results suggest that LSTM language models learn high-level representations that are related to representations in the human brain.

Why so gloomy? A Bayesian explanation of human pessimism bias in the multi-armed bandit task

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #100

Dalin Guo · Angela J Yu

How humans make repeated choices among options with imperfectly known reward outcomes is an important problem in psychology and neuroscience. This is often studied using multi-armed bandits, which is also frequently studied in machine learning. We present data from a human stationary bandit experiment, in which we vary the average abundance and variability of reward availability (mean and variance of reward rate distributions). Surprisingly, we find subjects significantly underestimate prior mean of reward rates -- based on their self-report, at the end of a game, on their reward expectation of non-chosen arms. Previously, human learning in the bandit task was found to be well captured by a Bayesian ideal learning model, the Dynamic Belief Model (DBM), albeit under an incorrect generative assumption of the temporal structure - humans assume reward rates can change over time even though they are actually fixed. We find that the "pessimism bias" in the bandit task is well captured by the prior mean of DBM when fitted to human choices; but it is poorly captured by the prior mean of the Fixed Belief Model (FBM), an alternative Bayesian model that (correctly) assumes reward rates to be constants. This pessimism bias is also incompletely captured by a simple reinforcement learning model (RL) commonly used in neuroscience and psychology, in terms of fitted initial Q-values. While it seems sub-optimal, and thus mysterious, that humans have an underestimated prior reward expectation, our simulations show that an underestimated prior mean helps to maximize long-term gain, if the observer assumes volatility when reward rates are stable and utilizes a softmax decision policy instead of the optimal one (obtainable by dynamic programming). This raises the intriguing possibility that the brain underestimates reward rates to compensate for the incorrect non-stationarity assumption in the generative model and a simplified decision policy.

Mental Sampling in Multimodal Representations

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #101

Jianqiao Zhu · Adam Sanborn · Nick Chater

Both resources in the natural environment and concepts in a semantic space are distributed "patchily", with large gaps in between the patches. To describe people's internal and external foraging behavior, various random walk models have been proposed. In particular, internal foraging has been modeled as sampling: in order to gather relevant information for making a decision, people draw samples from a mental representation using random-walk algorithms such as Markov chain Monte Carlo (MCMC). However, two common empirical observations argue against people using simple sampling algorithms such as MCMC for internal foraging. First, the distance between samples is often best described by a Levy flight distribution: the probability of the distance between two successive locations follows a power-law on the distances. Second, humans and other animals produce long-range, slowly decaying autocorrelations characterized as 1/f-like fluctuations, instead of the 1/f^2 fluctuations produced by random walks. We propose that mental sampling is not done by simple MCMC, but is instead adapted to multimodal representations and is implemented by Metropolis-coupled Markov chain Monte Carlo (MC3), one of the first algorithms developed for sampling from multimodal distributions. MC3 involves running multiple Markov chains in parallel but with target distributions of different temperatures, and it swaps the states of the chains whenever a better location is found. Heated chains more readily traverse valleys in the probability landscape to propose moves to far-away peaks, while the colder chains make the local steps that explore the current peak or patch. We show that MC3 generates distances between successive samples that follow a Levy flight distribution and produce 1/f-like autocorrelations, providing a single mechanistic account of these two puzzling empirical phenomena of internal foraging.

Integrated accounts of behavioral and neuroimaging data using flexible recurrent neural network models

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #102

Amir Dezfouli · Richard Morris · Fabio Ramos · Peter Dayan · Bernard Balleine

Neuroscience studies of human decision-making abilities commonly involve subjects completing a decision-making task while BOLD signals are recorded using fMRI. Hypotheses are tested about which brain regions mediate the effect of past experience, such as rewards, on future actions. One standard approach to this is model-based fMRI data analysis, in which a model is fitted to the behavioral data, i.e., a subject's choices, and then the neural data are parsed to find brain regions whose BOLD signals are related to the model's internal signals. However, the internal mechanics of such purely behavioral models are not constrained by the neural data, and therefore might miss or mischaracterize aspects of the brain. To address this limitation, we introduce a new method using recurrent neural network models that are flexible enough to be jointly fitted to the behavioral and neural data. We trained a model so that its internal states were suitably related to neural activity during the task, while at the same time its output predicted the next action a subject would execute. We then used the fitted model to create a novel visualization of the relationship between the activity in brain regions at different times following a reward and the choices the subject subsequently made. Finally, we validated our method using a previously published dataset. We found that the model was able to recover the underlying neural substrates that were discovered by explicit model engineering in the previous work, and also derived new results regarding the temporal pattern of brain activity.

Efficient inference for time-varying behavior during learning

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #103

Nicholas Roy · Ji Hyun Bak · Athena Akrami · Carlos Brody · Jonathan W Pillow

The process of learning new behaviors over time is a problem of great interest in both neuroscience and artificial intelligence. However, most standard analyses of animal training data either treat behavior as fixed or track only coarse performance statistics (e.g., accuracy, bias), providing limited insight into the evolution of the policies governing behavior. To overcome these limitations, we propose a dynamic psychophysical model that efficiently tracks trial-to-trial changes in behavior over the course of training. Our model consists of a dynamic logistic regression model, parametrized by a set of time-varying weights that express dependence on sensory stimuli as well as task-irrelevant covariates, such as stimulus, choice, and answer history. Our implementation scales to large behavioral datasets, allowing us to infer 500K parameters (e.g. 10 weights over 50K trials) in minutes on a desktop computer. We optimize hyperparameters governing how rapidly each weight evolves over time using the decoupled Laplace approximation, an efficient method for maximizing marginal likelihood in non-conjugate models. To illustrate performance, we apply our method to psychophysical data from both rats and human subjects learning a delayed sensory discrimination task. The model successfully tracks the psychophysical weights of rats over the course of training, capturing day-to-day and trial-to-trial fluctuations that underlie changes in performance, choice bias, and dependencies on task history. Finally, we investigate why rats frequently make mistakes on easy trials, and suggest that apparent lapses can be explained by sub-optimal weighting of known task covariates.

Multivariate Convolutional Sparse Coding for Electromagnetic Brain Signals

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #104

Tom Dupré la Tour · Thomas Moreau · Mainak Jas · Alexandre Gramfort

Frequency-specific patterns of neural activity are traditionally interpreted as sustained rhythmic oscillations, and related to cognitive mechanisms such as attention, high level visual processing or motor control. While alpha waves (8--12\,Hz) are known to closely resemble short sinusoids, and thus are revealed by Fourier analysis or wavelet transforms, there is an evolving debate that electromagnetic neural signals are composed of more complex waveforms that cannot be analyzed by linear filters and traditional signal representations. In this paper, we propose to learn dedicated representations of such recordings using a multivariate convolutional sparse coding (CSC) algorithm. Applied to electroencephalography (EEG) or magnetoencephalography (MEG) data, this method is able to learn not only prototypical temporal waveforms, but also associated spatial patterns so their origin can be localized in the brain. Our algorithm is based on alternated minimization and a greedy coordinate descent solver that leads to state-of-the-art running time on long time series. To demonstrate the implications of this method, we apply it to MEG data and show that it is able to recover biological artifacts. More remarkably, our approach also reveals the presence of non-sinusoidal mu-shaped patterns, along with their topographic maps related to the somatosensory cortex.

Manifold-tiling Localized Receptive Fields are Optimal in Similarity-preserving Neural Networks

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #105

Anirvan Sengupta · Cengiz Pehlevan · Mariano Tepper · Alexander Genkin · Dmitri Chklovskii

Many neurons in the brain, such as place cells in the rodent hippocampus, have localized receptive fields, i.e., they respond to a small neighborhood of stimulus space. What is the functional significance of such representations and how can they arise? Here, we propose that localized receptive fields emerge in similarity-preserving networks of rectifying neurons that learn low-dimensional manifolds populated by sensory inputs. Numerical simulations of such networks on standard datasets yield manifold-tiling localized receptive fields. More generally, we show analytically that, for data lying on symmetric manifolds, optimal solutions of objectives, from which similarity-preserving networks are derived, have localized receptive fields. Therefore, nonnegative similarity-preserving mapping (NSM) implemented by neural networks can model representations of continuous manifolds in the brain.

Connectionist Temporal Classification with Maximum Entropy Regularization

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #106

Hu Liu · Sheng Jin · Changshui Zhang

Connectionist Temporal Classification (CTC) is an objective function for end-to-end sequence learning, which adopts dynamic programming algorithms to directly learn the mapping between sequences. CTC has shown promising results in many sequence learning applications including speech recognition and scene text recognition. However, CTC tends to produce highly peaky and overconfident distributions, which is a symptom of overfitting. To remedy this, we propose a regularization method based on maximum conditional entropy which penalizes peaky distributions and encourages exploration. We also introduce an entropy-based pruning method to dramatically reduce the number of CTC feasible paths by ruling out unreasonable alignments. Experiments on scene text recognition show that our proposed methods consistently improve over the CTC baseline without the need to adjust training settings. Code has been made publicly available at: https://github.com/liuhu-bigeye/enctc.crnn.

Removing the Feature Correlation Effect of Multiplicative Noise

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #107

Zijun Zhang · Yining Zhang · Zongpeng Li

Multiplicative noise, including dropout, is widely used to regularize deep neural networks (DNNs), and is shown to be effective in a wide range of architectures and tasks. From an information perspective, we consider injecting multiplicative noise into a DNN as training the network to solve the task with noisy information pathways, which leads to the observation that multiplicative noise tends to increase the correlation between features, so as to increase the signal-to-noise ratio of information pathways. However, high feature correlation is undesirable, as it increases redundancy in representations. In this work, we propose non-correlating multiplicative noise (NCMN), which exploits batch normalization to remove the correlation effect in a simple yet effective way. We show that NCMN significantly improves the performance of standard multiplicative noise on image classification tasks, providing a better alternative to dropout for batch-normalized networks. Additionally, we present a unified view of NCMN and shake-shake regularization, which explains the performance gain of the latter.

Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #108

Mikhail Belkin · Daniel Hsu · Partha Mitra

Many modern machine learning models are trained to achieve zero or near-zero training error in order to obtain near-optimal (but non-zero) test error. This phenomenon of strong generalization performance for ``overfitted'' / interpolated classifiers appears to be ubiquitous in high-dimensional data, having been observed in deep networks, kernel machines, boosting and random forests. Their performance is consistently robust even when the data contain large amounts of label noise.

Very little theory is available to explain these observations. The vast majority of theoretical analyses of generalization allows for interpolation only when there is little or no label noise. This paper takes a step toward a theoretical foundation for interpolated classifiers by analyzing local interpolating schemes, including geometric simplicial interpolation algorithm and singularly weighted k-nearest neighbor schemes. Consistency or near-consistency is proved for these schemes in classification and regression problems. Moreover, the nearest neighbor schemes exhibit optimal rates under some standard statistical assumptions.

Finally, this paper suggests a way to explain the phenomenon of adversarial examples, which are seemingly ubiquitous in modern machine learning, and also discusses some connections to kernel machines and random forests in the interpolated regime.

Smoothed Analysis of Discrete Tensor Decomposition and Assemblies of Neurons

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #109

Nima Anari · Constantinos Daskalakis · Wolfgang Maass · Christos Papadimitriou · Amin Saberi · Santosh Vempala

We analyze linear independence of rank one tensors produced by tensor powers of randomly perturbed vectors. This enables efficient decomposition of sums of high-order tensors. Our analysis builds upon [BCMV14] but allows for a wider range of perturbation models, including discrete ones. We give an application to recovering assemblies of neurons. Assemblies are large sets of neurons representing specific memories or concepts. The size of the intersection of two assemblies has been shown in experiments to represent the extent to which these memories co-occur or these concepts are related; the phenomenon is called association of assemblies. This suggests that an animal's memory is a complex web of associations, and poses the problem of recovering this representation from cognitive data. Motivated by this problem, we study the following more general question: Can we reconstruct the Venn diagram of a family of sets, given the sizes of their l-wise intersections? We show that as long as the family of sets is randomly perturbed, it is enough for the number of measurements to be polynomially larger than the number of nonempty regions of the Venn diagram to fully reconstruct the diagram.

Entropy and mutual information in models of deep neural networks

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #110

Marylou Gabrié · Andre Manoel · Clément Luneau · jean barbier · Nicolas Macris · Florent Krzakala · Lenka Zdeborová

We examine a class of stochastic deep learning models with a tractable method to compute information-theoretic quantities. Our contributions are three-fold: (i) We show how entropies and mutual informations can be derived from heuristic statistical physics methods, under the assumption that weight matrices are independent and orthogonally-invariant. (ii) We extend particular cases in which this result is known to be rigorously exact by providing a proof for two-layers networks with Gaussian random weights, using the recently introduced adaptive interpolation method. (iii) We propose an experiment framework with generative models of synthetic datasets, on which we train deep neural networks with a weight constraint designed so that the assumption in (i) is verified during learning. We study the behavior of entropies and mutual information throughout learning and conclude that, in the proposed setting, the relationship between compression and generalization remains elusive.

The committee machine: Computational to statistical gaps in learning a two-layers neural network

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #111

Benjamin Aubin · Antoine Maillard · jean barbier · Florent Krzakala · Nicolas Macris · Lenka Zdeborová

Heuristic tools from statistical physics have been used in the past to compute the optimal learning and generalization errors in the teacher-student scenario in multi- layer neural networks. In this contribution, we provide a rigorous justification of these approaches for a two-layers neural network model called the committee machine. We also introduce a version of the approximate message passing (AMP) algorithm for the committee machine that allows to perform optimal learning in polynomial time for a large set of parameters. We find that there are regimes in which a low generalization error is information-theoretically achievable while the AMP algorithm fails to deliver it; strongly suggesting that no efficient algorithm exists for those cases, and unveiling a large computational gap.

A Unified Framework for Extensive-Form Game Abstraction with Bounds

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #112

Christian Kroer · Tuomas Sandholm

Abstraction has long been a key component in the practical solving of large-scale extensive-form games. Despite this, abstraction remains poorly understood. There have been some recent theoretical results but they have been confined to specific assumptions on abstraction structure and are specific to various disjoint types of abstraction, and specific solution concepts, for example, exact Nash equilibria or strategies with bounded immediate regret. In this paper we present a unified framework for analyzing abstractions that can express all types of abstractions and solution concepts used in prior papers with performance guarantees---while maintaining comparable bounds on abstraction quality. Moreover, our framework gives an exact decomposition of abstraction error in a much broader class of games, albeit only in an ex-post sense, as our results depend on the specific strategy chosen. Nonetheless, we use this ex-post decomposition along with slightly weaker assumptions than prior work to derive generalizations of prior bounds on abstraction quality. We also show, via counterexample, that such assumptions are necessary for some games. Finally, we prove the first bounds for how ϵ-Nash equilibria computed in abstractions perform in the original game. This is important because often one cannot afford to compute an exact Nash equilibrium in the abstraction. All our results apply to general-sum n-player games.

Connecting Optimization and Regularization Paths

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #113

Arun Suggala · Adarsh Prasad · Pradeep Ravikumar

We study the implicit regularization properties of optimization techniques by explicitly connecting their optimization paths to the regularization paths of ``corresponding'' regularized problems. This surprising connection shows that iterates of optimization techniques such as gradient descent and mirror descent are \emph{pointwise} close to solutions of appropriately regularized objectives. While such a tight connection between optimization and regularization is of independent intellectual interest, it also has important implications for machine learning: we can port results from regularized estimators to optimization, and vice versa. We investigate one key consequence, that borrows from the well-studied analysis of regularized estimators, to then obtain tight excess risk bounds of the iterates generated by optimization techniques.

Overlapping Clustering Models, and One (class) SVM to Bind Them All

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #114

Xueyu Mao · Purnamrita Sarkar · Deepayan Chakrabarti

People belong to multiple communities, words belong to multiple topics, and books cover multiple genres; overlapping clusters are commonplace. Many existing overlapping clustering methods model each person (or word, or book) as a non-negative weighted combination of "exemplars" who belong solely to one community, with some small noise. Geometrically, each person is a point on a cone whose corners are these exemplars. This basic form encompasses the widely used Mixed Membership Stochastic Blockmodel of networks and its degree-corrected variants, as well as topic models such as LDA. We show that a simple one-class SVM yields provably consistent parameter inference for all such models, and scales to large datasets. Experimental results on several simulated and real datasets show our algorithm (called SVM-cone) is both accurate and scalable.

Learning latent variable structured prediction models with Gaussian perturbations

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #115

Kevin Bello · Jean Honorio

The standard margin-based structured prediction commonly uses a maximum loss over all possible structured outputs. The large-margin formulation including latent variables not only results in a non-convex formulation but also increases the search space by a factor of the size of the latent space. Recent work has proposed the use of the maximum loss over random structured outputs sampled independently from some proposal distribution, with theoretical guarantees. We extend this work by including latent variables. We study a new family of loss functions under Gaussian perturbations and analyze the effect of the latent space on the generalization bounds. We show that the non-convexity of learning with latent variables originates naturally, as it relates to a tight upper bound of the Gibbs decoder distortion with respect to the latent space. Finally, we provide a formulation using random samples and relaxations that produces a tighter upper bound of the Gibbs decoder distortion up to a statistical accuracy, which enables a polynomial time evaluation of the objective function. We illustrate the method with synthetic experiments and a computer vision application.

Self-Supervised Generation of Spatial Audio for 360° Video

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #116

Pedro Morgado · Nuno Nvasconcelos · Timothy Langlois · Oliver Wang

We introduce an approach to convert mono audio recorded by a 360° video camera into spatial audio, a representation of the distribution of sound over the full viewing sphere. Spatial audio is an important component of immersive 360° video viewing, but spatial audio microphones are still rare in current 360° video production. Our system consists of end-to-end trainable neural networks that separate individual sound sources and localize them on the viewing sphere, conditioned on multi-modal analysis from the audio and 360° video frames. We introduce several datasets, including one filmed ourselves, and one collected in-the-wild from YouTube, consisting of 360° videos uploaded with spatial audio. During training, ground truth spatial audio serves as self-supervision and a mixed down mono track forms the input to our network. Using our approach we show that it is possible to infer the spatial localization of sounds based only on a synchronized 360° video and the mono audio track.

Symbolic Graph Reasoning Meets Convolutions

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #117

Xiaodan Liang · Zhiting Hu · Hao Zhang · Liang Lin · Eric Xing

Beyond local convolution networks, we explore how to harness various external human knowledge for endowing the networks with the capability of semantic global reasoning. Rather than using separate graphical models (e.g. CRF) or constraints for modeling broader dependencies, we propose a new Symbolic Graph Reasoning (SGR) layer, which performs reasoning over a group of symbolic nodes whose outputs explicitly represent different properties of each semantic in a prior knowledge graph. To cooperate with local convolutions, each SGR is constituted by three modules: a) a primal local-to-semantic voting module where the features of all symbolic nodes are generated by voting from local representations; b) a graph reasoning module propagates information over knowledge graph to achieve global semantic coherency; c) a dual semantic-to-local mapping module learns new associations of the evolved symbolic nodes with local representations, and accordingly enhances local features. The SGR layer can be injected between any convolution layers and instantiated with distinct prior graphs. Extensive experiments show incorporating SGR significantly improves plain ConvNets on three semantic segmentation tasks and one image classification task. More analyses show the SGR layer learns shared symbolic representations for domains/datasets with the different label set given a universal knowledge graph, demonstrating its superior generalization capability.

Towards Deep Conversational Recommendations

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #118

Raymond Li · Samira Ebrahimi Kahou · Hannes Schulz · Vincent Michalski · Laurent Charlin · Chris Pal

There has been growing interest in using neural networks and deep learning techniques to create dialogue systems. Conversational recommendation is an interesting setting for the scientific exploration of dialogue with natural language as the associated discourse involves goal-driven dialogue that often transforms naturally into more free-form chat. This paper provides two contributions. First, until now there has been no publicly available large-scale data set consisting of real-world dialogues centered around recommendations. To address this issue and to facilitate our exploration here, we have collected ReDial, a data set consisting of over 10,000 conversations centered around the theme of providing movie recommendations. We make this data available to the community for further research. Second, we use this dataset to explore multiple facets of conversational recommendations. In particular we explore new neural architectures, mechanisms and methods suitable for composing conversational recommendation systems. Our dataset allows us to systematically probe model sub-components addressing different parts of the overall problem domain ranging from: sentiment analysis and cold-start recommendation generation to detailed aspects of how natural language is used in this setting in the real world. We combine such sub-components into a full-blown dialogue system and examine its behavior.

Human-in-the-Loop Interpretability Prior

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #119

Isaac Lage · Andrew Ross · Samuel J Gershman · Been Kim · Finale Doshi-Velez

We often desire our models to be interpretable as well as accurate. Prior work on optimizing models for interpretability has relied on easy-to-quantify proxies for interpretability, such as sparsity or the number of operations required. In this work, we optimize for interpretability by directly including humans in the optimization loop. We develop an algorithm that minimizes the number of user studies to find models that are both predictive and interpretable and demonstrate our approach on several data sets. Our human subjects results show trends towards different proxy notions of interpretability on different datasets, which suggests that different proxies are preferred on different tasks.

Why Is My Classifier Discriminatory?

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #120

Irene Chen · Fredrik Johansson · David Sontag

Recent attempts to achieve fairness in predictive models focus on the balance between fairness and accuracy. In sensitive applications such as healthcare or criminal justice, this trade-off is often undesirable as any increase in prediction error could have devastating consequences. In this work, we argue that the fairness of predictions should be evaluated in context of the data, and that unfairness induced by inadequate samples sizes or unmeasured predictive variables should be addressed through data collection, rather than by constraining the model. We decompose cost-based metrics of discrimination into bias, variance, and noise, and propose actions aimed at estimating and reducing each term. Finally, we perform case-studies on prediction of income, mortality, and review ratings, confirming the value of this analysis. We find that data collection is often a means to reduce discrimination without sacrificing accuracy.

Link Prediction Based on Graph Neural Networks

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #121

Muhan Zhang · Yixin Chen

Link prediction is a key problem for network-structured data. Link prediction heuristics use some score functions, such as common neighbors and Katz index, to measure the likelihood of links. They have obtained wide practical uses due to their simplicity, interpretability, and for some of them, scalability. However, every heuristic has a strong assumption on when two nodes are likely to link, which limits their effectiveness on networks where these assumptions fail. In this regard, a more reasonable way should be learning a suitable heuristic from a given network instead of using predefined ones. By extracting a local subgraph around each target link, we aim to learn a function mapping the subgraph patterns to link existence, thus automatically learning a ``heuristic'' that suits the current network. In this paper, we study this heuristic learning paradigm for link prediction. First, we develop a novel γ-decaying heuristic theory. The theory unifies a wide range of heuristics in a single framework, and proves that all these heuristics can be well approximated from local subgraphs. Our results show that local subgraphs reserve rich information related to link existence. Second, based on the γ-decaying theory, we propose a new method to learn heuristics from local subgraphs using a graph neural network (GNN). Its experimental results show unprecedented performance, working consistently well on a wide range of problems.

KONG: Kernels for ordered-neighborhood graphs

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #122

Moez Draief · Konstantin Kutzkov · Kevin Scaman · Milan Vojnovic

We present novel graph kernels for graphs with node and edge labels that have ordered neighborhoods, i.e. when neighbor nodes follow an order. Graphs with ordered neighborhoods are a natural data representation for evolving graphs where edges are created over time, which induces an order. Combining convolutional subgraph kernels and string kernels, we design new scalable algorithms for generation of explicit graph feature maps using sketching techniques. We obtain precise bounds for the approximation accuracy and computational complexity of the proposed approaches and demonstrate their applicability on real datasets. In particular, our experiments demonstrate that neighborhood ordering results in more informative features. For the special case of general graphs, i.e. graphs without ordered neighborhoods, the new graph kernels yield efficient and simple algorithms for the comparison of label distributions between graphs.

Efficient Stochastic Gradient Hard Thresholding

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #123

Pan Zhou · Xiaotong Yuan · Jiashi Feng

Stochastic gradient hard thresholding methods have recently been shown to work favorably in solving large-scale empirical risk minimization problems under sparsity or rank constraint. Despite the improved iteration complexity over full gradient methods, the gradient evaluation and hard thresholding complexity of the existing stochastic algorithms usually scales linearly with data size, which could still be expensive when data is huge and the hard thresholding step could be as expensive as singular value decomposition in rank-constrained problems. To address these deficiencies, we propose an efficient hybrid stochastic gradient hard thresholding (HSG-HT) method that can be provably shown to have sample-size-independent gradient evaluation and hard thresholding complexity bounds. Specifically, we prove that the stochastic gradient evaluation complexity of HSG-HT scales linearly with inverse of sub-optimality and its hard thresholding complexity scales logarithmically. By applying the heavy ball acceleration technique, we further propose an accelerated variant of HSG-HT which can be shown to have improved factor dependence on restricted condition number. Numerical results confirm our theoretical affirmation and demonstrate the computational efficiency of the proposed methods.

Measures of distortion for machine learning

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #124

Leena Chennuru Vankadara · Ulrike von Luxburg

Given data from a general metric space, one of the standard machine learning pipelines is to first embed the data into a Euclidean space and subsequently apply out of the box machine learning algorithms to analyze the data. The quality of such an embedding is typically described in terms of a distortion measure. In this paper, we show that many of the existing distortion measures behave in an undesired way, when considered from a machine learning point of view. We investigate desirable properties of distortion measures and formally prove that most of the existing measures fail to satisfy these properties. These theoretical findings are supported by simulations, which for example demonstrate that existing distortion measures are not robust to noise or outliers and cannot serve as good indicators for classification accuracy. As an alternative, we suggest a new measure of distortion, called σ-distortion. We can show both in theory and in experiments that it satisfies all desirable properties and is a better candidate to evaluate distortion in the context of machine learning.

Relating Leverage Scores and Density using Regularized Christoffel Functions

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #125

Edouard Pauwels · Francis Bach · Jean-Philippe Vert

Statistical leverage scores emerged as a fundamental tool for matrix sketching and column sampling with applications to low rank approximation, regression, random feature learning and quadrature. Yet, the very nature of this quantity is barely understood. Borrowing ideas from the orthogonal polynomial literature, we introduce the regularized Christoffel function associated to a positive definite kernel. This uncovers a variational formulation for leverage scores for kernel methods and allows to elucidate their relationships with the chosen kernel as well as population density. Our main result quantitatively describes a decreasing relation between leverage score and population density for a broad class of kernels on Euclidean spaces. Numerical simulations support our findings.

Streaming Kernel PCA with $\tilde{O}(\sqrt{n})$ Random Features

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #126

Md Enayat Ullah · Poorya Mianjy · Teodor Vanislavov Marinov · Raman Arora

We study the statistical and computational aspects of kernel principal component analysis using random Fourier features and show that under mild assumptions, $O(\sqrt{n} \log n)$ features suffices to achieve $O(1/\epsilon^2)$ sample complexity. Furthermore, we give a memory efficient streaming algorithm based on classical Oja's algorithm that achieves this rate

---

Learning with SGD and Random Features

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #127

Luigi Carratino · Alessandro Rudi · Lorenzo Rosasco

Sketching and stochastic gradient methods are arguably the most common techniques to derive efficient large scale learning algorithms. In this paper, we investigate their application in the context of nonparametric statistical learning. More precisely, we study the estimator defined by stochastic gradient with mini batches and random features. The latter can be seen as form of nonlinear sketching and used to define approximate kernel methods. The considered estimator is not explicitly penalized/constrained and regularization is implicit. Indeed, our study highlights how different parameters, such as number of features, iterations, step-size and mini-batch size control the learning properties of the solutions. We do this by deriving optimal finite sample bounds, under standard assumptions. The obtained results are corroborated and illustrated by numerical experiments.

But How Does It Work in Theory? Linear SVM with Random Features

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #128

Yitong Sun · Anna Gilbert · Ambuj Tewari

We prove that, under low noise assumptions, the support vector machine with $N\ll m$ random features (RFSVM) can achieve the learning rate faster than $O(1/\sqrt{m})$ on a training set with m samples when an optimized feature map is used. Our work extends the previous fast rate analysis of random features method from least square loss to 0-1 loss. We also show that the reweighted feature selection method, which approximates the optimized feature map, helps improve the performance of RFSVM in experiments on a synthetic data set.

Statistical and Computational Trade-Offs in Kernel K-Means

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #129

Daniele Calandriello · Lorenzo Rosasco

We investigate the efficiency of k-means in terms of both statistical and computational requirements. More precisely, we study a Nystr\"om approach to kernel k-means. We analyze the statistical properties of the proposed method and show that it achieves the same accuracy of exact kernel k-means with only a fraction of computations. Indeed, we prove under basic assumptions that sampling $\sqrt{n}$ Nystr\"om landmarks allows to greatly reduce computational costs without incurring in any loss of accuracy. To the best of our knowledge this is the first result showing in this kind for unsupervised learning.

Quadrature-based features for kernel approximation

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #130

Marina Munkhoeva · Yermek Kapushev · Evgeny Burnaev · Ivan Oseledets

We consider the problem of improving kernel approximation via randomized feature maps. These maps arise as Monte Carlo approximation to integral representations of kernel functions and scale up kernel methods for larger datasets. Based on an efficient numerical integration technique, we propose a unifying approach that reinterprets the previous random features methods and extends to better estimates of the kernel approximation. We derive the convergence behavior and conduct an extensive empirical study that supports our hypothesis.

Processing of missing data by neural networks

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #131

Marek Śmieja · Łukasz Struski · Jacek Tabor · Bartosz Zieliński · Przemysław Spurek

We propose a general, theoretically justified mechanism for processing missing data by neural networks. Our idea is to replace typical neuron's response in the first hidden layer by its expected value. This approach can be applied for various types of networks at minimal cost in their modification. Moreover, in contrast to recent approaches, it does not require complete data for training. Experimental results performed on different types of architectures show that our method gives better results than typical imputation strategies and other methods dedicated for incomplete data.

Constructing Deep Neural Networks by Bayesian Network Structure Learning

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #132

Raanan Y. Rohekar · Shami Nisimov · Yaniv Gurwicz · Guy Koren · Gal Novik

We introduce a principled approach for unsupervised structure learning of deep neural networks. We propose a new interpretation for depth and inter-layer connectivity where conditional independencies in the input distribution are encoded hierarchically in the network structure. Thus, the depth of the network is determined inherently. The proposed method casts the problem of neural network structure learning as a problem of Bayesian network structure learning. Then, instead of directly learning the discriminative structure, it learns a generative graph, constructs its stochastic inverse, and then constructs a discriminative graph. We prove that conditional-dependency relations among the latent variables in the generative graph are preserved in the class-conditional discriminative graph. We demonstrate on image classification benchmarks that the deepest layers (convolutional and dense) of common networks can be replaced by significantly smaller learned structures, while maintaining classification accuracy---state-of-the-art on tested benchmarks. Our structure learning algorithm requires a small computational cost and runs efficiently on a standard desktop CPU.

Mallows Models for Top-k Lists

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #133

Flavio Chierichetti · Anirban Dasgupta · Shahrzad Haddadan · Ravi Kumar · Silvio Lattanzi

The classic Mallows model is a widely-used tool to realize distributions on per- mutations. Motivated by common practical situations, in this paper, we generalize Mallows to model distributions on top-k lists by using a suitable distance measure between top-k lists. Unlike many earlier works, our model is both analytically tractable and computationally efficient. We demonstrate this by studying two basic problems in this model, namely, sampling and reconstruction, from both algorithmic and experimental points of view.

Cooperative neural networks (CoNN): Exploiting prior independence structure for improved classification

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #134

Harsh Shrivastava · Eugene Bart · Bob Price · Hanjun Dai · Bo Dai · Srinivas Aluru

We propose a new approach, called cooperative neural networks (CoNN), which use a set of cooperatively trained neural networks to capture latent representations that exploit prior given independence structure. The model is more flexible than traditional graphical models based on exponential family distributions, but incorporates more domain specific prior structure than traditional deep networks or variational autoencoders. The framework is very general and can be used to exploit the independence structure of any graphical model. We illustrate the technique by showing that we can transfer the independence structure of the popular Latent Dirichlet Allocation

(LDA) model to a cooperative neural network, CoNN-sLDA. Empirical evaluation of CoNN-sLDA on supervised text classification tasks demonstrate that the theoretical advantages of prior independence structure can be realized in practice - we demonstrate a 23 percent reduction in error on the challenging MultiSent data set compared to state-of-the-art.

Maximum-Entropy Fine Grained Classification

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #135

Abhimanyu Dubey · Otkrist Gupta · Ramesh Raskar · Nikhil Naik

Fine-Grained Visual Classification (FGVC) is an important computer vision problem that involves small diversity within the different classes, and often requires expert annotators to collect data. Utilizing this notion of small visual diversity, we revisit Maximum-Entropy learning in the context of fine-grained classification, and provide a training routine that maximizes the entropy of the output probability distribution for training convolutional neural networks on FGVC tasks. We provide a theoretical as well as empirical justification of our approach, and achieve state-of-the-art performance across a variety of classification tasks in FGVC, that can potentially be extended to any fine-tuning task. Our method is robust to different hyperparameter values, amount of training data and amount of training label noise and can hence be a valuable tool in many similar problems.

Efficient Loss-Based Decoding on Graphs for Extreme Classification

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #136

Itay Evron · Edward Moroshko · Koby Crammer

In extreme classification problems, learning algorithms are required to map instances to labels from an extremely large label set. We build on a recent extreme classification framework with logarithmic time and space (LTLS), and on a general approach for error correcting output coding (ECOC) with loss-based decoding, and introduce a flexible and efficient approach accompanied by theoretical bounds. Our framework employs output codes induced by graphs, for which we show how to perform efficient loss-based decoding to potentially improve accuracy. In addition, our framework offers a tradeoff between accuracy, model size and prediction time. We show how to find the sweet spot of this tradeoff using only the training data. Our experimental study demonstrates the validity of our assumptions and claims, and shows that our method is competitive with state-of-the-art algorithms.

A no-regret generalization of hierarchical softmax to extreme multi-label classification

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #137

Marek Wydmuch · Kalina Jasinska · Mikhail Kuznetsov · Róbert Busa-Fekete · Krzysztof Dembczynski

Extreme multi-label classification (XMLC) is a problem of tagging an instance with a small subset of relevant labels chosen from an extremely large pool of possible labels. Large label spaces can be efficiently handled by organizing labels as a tree, like in the hierarchical softmax (HSM) approach commonly used for multi-class problems. In this paper, we investigate probabilistic label trees (PLTs) that have been recently devised for tackling XMLC problems. We show that PLTs are a no-regret multi-label generalization of HSM when precision@k is used as a model evaluation metric. Critically, we prove that pick-one-label heuristic---a reduction technique from multi-label to multi-class that is routinely used along with HSM---is not consistent in general. We also show that our implementation of PLTs, referred to as extremeText (XT), obtains significantly better results than HSM with the pick-one-label heuristic and XML-CNN, a deep network specifically designed for XMLC problems. Moreover, XT is competitive to many state-of-the-art approaches in terms of statistical performance, model size and prediction time which makes it amenable to deploy in an online system.

Efficient Gradient Computation for Structured Output Learning with Rational and Tropical Losses

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #138

Corinna Cortes · Vitaly Kuznetsov · Mehryar Mohri · Dmitry Storcheus · Scott Yang

Many structured prediction problems admit a natural loss function for evaluation such as the edit-distance or n-gram loss. However, existing learning algorithms are typically designed to optimize alternative objectives such as the cross-entropy. This is because a na\"{i}ve implementation of the natural loss functions often results in intractable gradient computations. In this paper, we design efficient gradient computation algorithms for two broad families of structured prediction loss functions: rational and tropical losses. These families include as special cases the n-gram loss, the edit-distance loss, and many other loss functions commonly used in natural language processing and computational biology tasks that are based on sequence similarity measures. Our algorithms make use of weighted automata and graph operations over appropriate semirings to design efficient solutions. They facilitate efficient gradient computation and hence enable one to train learning models such as neural networks with complex structured losses.

Deep Structured Prediction with Nonlinear Output Transformations

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #139

Colin Graber · Ofer Meshi · Alexander Schwing

Deep structured models are widely used for tasks like semantic segmentation, where explicit correlations between variables provide important prior information which generally helps to reduce the data needs of deep nets. However, current deep structured models are restricted by oftentimes very local neighborhood structure, which cannot be increased for computational complexity reasons, and by the fact that the output configuration, or a representation thereof, cannot be transformed further. Very recent approaches which address those issues include graphical model inference inside deep nets so as to permit subsequent non-linear output space transformations. However, optimization of those formulations is challenging and not well understood. Here, we develop a novel model which generalizes existing approaches, such as structured prediction energy networks, and discuss a formulation which maintains applicability of existing inference techniques.

Mapping Images to Scene Graphs with Permutation-Invariant Structured Prediction

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #140

Roei Herzig · Moshiko Raboh · Gal Chechik · Jonathan Berant · Amir Globerson

Machine understanding of complex images is a key goal of artificial intelligence. One challenge underlying this task is that visual scenes contain multiple inter-related objects, and that global context plays an important role in interpreting the scene. A natural modeling framework for capturing such effects is structured prediction, which optimizes over complex labels, while modeling within-label interactions. However, it is unclear what principles should guide the design of a structured prediction model that utilizes the power of deep learning components. Here we propose a design principle for such architectures that follows from a natural requirement of permutation invariance. We prove a necessary and sufficient characterization for architectures that follow this invariance, and discuss its implication on model design. Finally, we show that the resulting model achieves new state of the art results on the Visual Genome scene graph labeling benchmark, outperforming all recent approaches.

Large Margin Deep Networks for Classification

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #141

Gamaleldin Elsayed · Dilip Krishnan · Hossein Mobahi · Kevin Regan · Samy Bengio

We present a formulation of deep learning that aims at producing a large margin classifier. The notion of \emc{margin}, minimum distance to a decision boundary, has served as the foundation of several theoretically profound and empirically successful results for both classification and regression tasks. However, most large margin algorithms are applicable only to shallow models with a preset feature representation; and conventional margin methods for neural networks only enforce margin at the output layer. Such methods are therefore not well suited for deep networks. In this work, we propose a novel loss function to impose a margin on any chosen set of layers of a deep network (including input and hidden layers). Our formulation allows choosing any \$l_p\$ norm (\$p \geq 1\$) on the metric measuring the margin. We demonstrate that the decision boundary obtained by our loss has nice properties compared to standard classification loss functions. Specifically, we show improved empirical results on the MNIST, CIFAR-10 and ImageNet datasets on multiple tasks: generalization from small training sets, corrupted labels, and robustness against adversarial perturbations. The resulting loss is general and complementary to existing data augmentation (such as random/adversarial input transform) and regularization techniques such as weight decay, dropout, and batch norm. \footnote{Code for the large margin loss function is released at \url{https://github.com/google-research/google-research/tree/master/large_margin}}

Semi-supervised Deep Kernel Learning: Regression with Unlabeled Data by Minimizing Predictive Variance

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #142

Neal Jean · Sang Michael Xie · Stefano Ermon

Large amounts of labeled data are typically required to train deep learning models. For many real-world problems, however, acquiring additional data can be expensive or even impossible. We present semi-supervised deep kernel learning (SSDKL), a semi-supervised regression model based on minimizing predictive variance in the posterior regularization framework. SSDKL combines the hierarchical representation learning of neural networks with the probabilistic modeling capabilities of Gaussian processes. By leveraging unlabeled data, we show improvements on a diverse set of real-world regression tasks over supervised deep kernel learning and semi-supervised methods such as VAT and mean teacher adapted for regression.

Multitask Boosting for Survival Analysis with Competing Risks

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #143

Alexis Bellot · Mihaela van der Schaar

The co-occurrence of multiple diseases among the general population is an important problem as those patients have more risk of complications and represent a large share of health care expenditure. Learning to predict time-to-event probabilities for these patients is a challenging problem because the risks of events are correlated (there are competing risks) with often only few patients experiencing individual events of interest, and of those only a fraction are actually observed in the data. We introduce in this paper a survival model with the flexibility to leverage a common representation of related events that is designed to correct for the strong imbalance in observed outcomes. The procedure is sequential: outcome-specific survival distributions form the components of nonparametric multivariate estimators which we combine into an ensemble in such a way as to ensure accurate predictions on all outcome types simultaneously. Our algorithm is general and represents the first boosting-like method for time-to-event data with multiple outcomes. We demonstrate the performance of our algorithm on synthetic and real data.

Multi-Layered Gradient Boosting Decision Trees

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #144

Ji Feng · Yang Yu · Zhi-Hua Zhou

Multi-layered distributed representation is believed to be the key ingredient of deep neural networks especially in cognitive tasks like computer vision. While non-differentiable models such as gradient boosting decision trees (GBDTs) are still the dominant methods for modeling discrete or tabular data, they are hard to incorporate with such representation learning ability. In this work, we propose the multi-layered GBDT forest (mGBDTs), with an explicit emphasis on exploring the ability to learn hierarchical distributed representations by stacking several layers of regression GBDTs as its building block. The model can be jointly trained by a variant of target propagation across layers, without the need to derive backpropagation nor differentiability. Experiments confirmed the effectiveness of the model in terms of performance and representation learning ability.

Unsupervised Adversarial Invariance

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #145

Ayush Jaiswal · Rex Yue Wu · Wael Abd-Almageed · Prem Natarajan

Data representations that contain all the information about target variables but are invariant to nuisance factors benefit supervised learning algorithms by preventing them from learning associations between these factors and the targets, thus reducing overfitting. We present a novel unsupervised invariance induction framework for neural networks that learns a split representation of data through competitive training between the prediction task and a reconstruction task coupled with disentanglement, without needing any labeled information about nuisance factors or domain knowledge. We describe an adversarial instantiation of this framework and provide analysis of its working. Our unsupervised model outperforms state-of-the-art methods, which are supervised, at inducing invariance to inherent nuisance factors, effectively using synthetic data augmentation to learn invariance, and domain adaptation. Our method can be applied to any prediction task, eg., binary/multi-class classification or regression, without loss of generality.

Learning Deep Disentangled Embeddings With the F-Statistic Loss

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #146

Karl Ridgeway · Michael Mozer

Deep-embedding methods aim to discover representations of a domain that make explicit the domain's class structure and thereby support few-shot learning. Disentangling methods aim to make explicit compositional or factorial structure. We combine these two active but independent lines of research and propose a new paradigm suitable for both goals. We propose and evaluate a novel loss function based on the F statistic, which describes the separation of two or more distributions. By ensuring that distinct classes are well separated on a subset of embedding dimensions, we obtain embeddings that are useful for few-shot learning. By not requiring separation on all dimensions, we encourage the discovery of disentangled representations. Our embedding method matches or beats state-of-the-art, as evaluated by performance on recall@k and few-shot learning tasks. Our method also obtains performance superior to a variety of alternatives on disentangling, as evaluated by two key properties of a disentangled representation: modularity and explicitness. The goal of our work is to obtain more interpretable, manipulable, and generalizable deep representations of concepts and categories.

Learning Latent Subspaces in Variational Autoencoders

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #147

Jack Klys · Jake Snell · Richard Zemel

Variational autoencoders (VAEs) are widely used deep generative models capable of learning unsupervised latent representations of data. Such representations are often difficult to interpret or control. We consider the problem of unsupervised learning of features correlated to specific labels in a dataset. We propose a VAE-based generative model which we show is capable of extracting features correlated to binary labels in the data and structuring it in a latent subspace which is easy to interpret. Our model, the Conditional Subspace VAE (CSVAE), uses mutual information minimization to learn a low-dimensional latent subspace associated with each label that can easily be inspected and independently manipulated. We demonstrate the utility of the learned representations for attribute manipulation tasks on both the Toronto Face and CelebA datasets.

Dual Swap Disentangling

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #148

Zunlei Feng · Xinchao Wang · Chenglong Ke · An-Xiang Zeng · Dacheng Tao · Mingli Song

Learning interpretable disentangled representations is a crucial yet challenging task. In this paper, we propose a weakly semi-supervised method, termed as Dual Swap Disentangling (DSD), for disentangling using both labeled and unlabeled data. Unlike conventional weakly supervised methods that rely on full annotations on the group of samples, we require only limited annotations on paired samples that indicate their shared attribute like the color. Our model takes the form of a dual autoencoder structure. To achieve disentangling using the labeled pairs, we follow a encoding-swap-decoding'' process, where we first swap the parts of their encodings corresponding to the shared attribute, and then decode the obtained hybrid codes to reconstruct the original input pairs. For unlabeled pairs, we follow theencoding-swap-decoding'' process twice on designated encoding parts and enforce the final outputs to approximate the input pairs. By isolating parts of the encoding and swapping them back and forth, we impose the dimension-wise modularity and portability of the encodings of the unlabeled samples, which implicitly encourages disentangling under the guidance of labeled pairs. This dual swap mechanism, tailored for semi-supervised setting, turns out to be very effective. Experiments on image datasets from a wide domain show that our model yields state-of-the-art disentangling performances.

Joint Autoregressive and Hierarchical Priors for Learned Image Compression

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #149

David Minnen · Johannes Ballé · Johannes Ballé · George D Toderici

Recent models for learned image compression are based on autoencoders that learn approximately invertible mappings from pixels to a quantized latent representation. The transforms are combined with an entropy model, which is a prior on the latent representation that can be used with standard arithmetic coding algorithms to generate a compressed bitstream. Recently, hierarchical entropy models were introduced as a way to exploit more structure in the latents than previous fully factorized priors, improving compression performance while maintaining end-to-end optimization. Inspired by the success of autoregressive priors in probabilistic generative models, we examine autoregressive, hierarchical, and combined priors as alternatives, weighing their costs and benefits in the context of image compression. While it is well known that autoregressive models can incur a significant computational penalty, we find that in terms of compression performance, autoregressive and hierarchical priors are complementary and can be combined to exploit the probabilistic structure in the latents better than all previous learned models. The combined model yields state-of-the-art rate-distortion performance and generates smaller files than existing methods: 15.8% rate reductions over the baseline hierarchical model and 59.8%, 35%, and 8.4% savings over JPEG, JPEG2000, and BPG, respectively. To the best of our knowledge, our model is the first learning-based method to outperform the top standard image codec (BPG) on both the PSNR and MS-SSIM distortion metrics.

Group Equivariant Capsule Networks

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #150

Jan Eric Lenssen · Matthias Fey · Pascal Libuschewski

We present group equivariant capsule networks, a framework to introduce guaranteed equivariance and invariance properties to the capsule network idea. Our work can be divided into two contributions. First, we present a generic routing by agreement algorithm defined on elements of a group and prove that equivariance of output pose vectors, as well as invariance of output activations, hold under certain conditions. Second, we connect the resulting equivariant capsule networks with work from the field of group convolutional networks. Through this connection, we provide intuitions of how both methods relate and are able to combine the strengths of both approaches in one deep neural network architecture. The resulting framework allows sparse evaluation of the group convolution operator, provides control over specific equivariance and invariance properties, and can use routing by agreement instead of pooling operations. In addition, it is able to provide interpretable and equivariant representation vectors as output capsules, which disentangle evidence of object existence from its pose.

Learning Disentangled Joint Continuous and Discrete Representations

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #151

Emilien Dupont

We present a framework for learning disentangled and interpretable jointly continuous and discrete representations in an unsupervised manner. By augmenting the continuous latent distribution of variational autoencoders with a relaxed discrete distribution and controlling the amount of information encoded in each latent unit, we show how continuous and categorical factors of variation can be discovered automatically from data. Experiments show that the framework disentangles continuous and discrete generative factors on various datasets and outperforms current disentangling methods when a discrete generative factor is prominent.

Image-to-image translation for cross-domain disentanglement

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #152

Abel Gonzalez-Garcia · Joost van de Weijer · Yoshua Bengio

Deep image translation methods have recently shown excellent results, outputting high-quality images covering multiple modes of the data distribution. There has also been increased interest in disentangling the internal representations learned by deep methods to further improve their performance and achieve a finer control. In this paper, we bridge these two objectives and introduce the concept of cross-domain disentanglement. We aim to separate the internal representation into three parts. The shared part contains information for both domains. The exclusive parts, on the other hand, contain only factors of variation that are particular to each domain. We achieve this through bidirectional image translation based on Generative Adversarial Networks and cross-domain autoencoders, a novel network component. Our model offers multiple advantages. We can output diverse samples covering multiple modes of the distributions of both domains, perform domain-specific image transfer and interpolation, and cross-domain retrieval without the need of labeled data, only paired images. We compare our model to the state-of-the-art in multi-modal image translation and achieve better results for translation on challenging datasets as well as for cross-domain retrieval on realistic datasets.

Cooperative Learning of Audio and Video Models from Self-Supervised Synchronization

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #153

Bruno Korbar · Du Tran · Lorenzo Torresani

There is a natural correlation between the visual and auditive elements of a video. In this work we leverage this connection to learn general and effective models for both audio and video analysis from self-supervised temporal synchronization. We demonstrate that a calibrated curriculum learning scheme, a careful choice of negative examples, and the use of a contrastive loss are critical ingredients to obtain powerful multi-sensory representations from models optimized to discern temporal synchronization of audio-video pairs. Without further fine-tuning, the resulting audio features achieve performance superior or comparable to the state-of-the-art on established audio classification benchmarks (DCASE2014 and ESC-50). At the same time, our visual subnet provides a very effective initialization to improve the accuracy of video-based action recognition models: compared to learning from scratch, our self-supervised pretraining yields a remarkable gain of +19.9% in action recognition accuracy on UCF101 and a boost of +17.7% on HMDB51.

Non-Adversarial Mapping with VAEs

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #154

Yedid Hoshen

The study of cross-domain mapping without supervision has recently attracted much attention. Much of the recent progress was enabled by the use of adversarial training as well as cycle constraints. The practical difficulty of adversarial training motivates research into non-adversarial methods. In a recent paper, it was shown that cross-domain mapping is possible without the use of cycles or GANs. Although promising, this approach suffers from several drawbacks including costly inference and an optimization variable for every training example preventing the method from using large training sets. We present an alternative approach which is able to achieve non-adversarial mapping using a novel form of Variational Auto-Encoder. Our method is much faster at inference time, is able to leverage large datasets and has a simple interpretation.

Learning to Teach with Dynamic Loss Functions

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #155

Lijun Wu · Fei Tian · Yingce Xia · Yang Fan · Tao Qin · Lai Jian-Huang · Tie-Yan Liu

Teaching is critical to human society: it is with teaching that prospective students are educated and human civilization can be inherited and advanced. A good teacher not only provides his/her students with qualified teaching materials (e.g., textbooks), but also sets up appropriate learning objectives (e.g., course projects and exams) considering different situations of a student. When it comes to artificial intelligence, treating machine learning models as students, the loss functions that are optimized act as perfect counterparts of the learning objective set by the teacher. In this work, we explore the possibility of imitating human teaching behaviors by dynamically and automatically outputting appropriate loss functions to train machine learning models. Different from typical learning settings in which the loss function of a machine learning model is predefined and fixed, in our framework, the loss function of a machine learning model (we call it student) is defined by another machine learning model (we call it teacher). The ultimate goal of teacher model is cultivating the student to have better performance measured on development dataset. Towards that end, similar to human teaching, the teacher, a parametric model, dynamically outputs different loss functions that will be used and optimized by its student model at different training stages. We develop an efficient learning method for the teacher model that makes gradient based optimization possible, exempt of the ineffective solutions such as policy optimization. We name our method as ``learning to teach with dynamic loss functions'' (L2T-DLF for short). Extensive experiments on real world tasks including image classification and neural machine translation demonstrate that our method significantly improves the quality of various student models.

Maximizing acquisition functions for Bayesian optimization

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #156

James Wilson · Frank Hutter · Marc Deisenroth

Bayesian optimization is a sample-efficient approach to global optimization that relies on theoretically motivated value heuristics (acquisition functions) to guide its search process. Fully maximizing acquisition functions produces the Bayes' decision rule, but this ideal is difficult to achieve since these functions are frequently non-trivial to optimize. This statement is especially true when evaluating queries in parallel, where acquisition functions are routinely non-convex, high-dimensional, and intractable. We first show that acquisition functions estimated via Monte Carlo integration are consistently amenable to gradient-based optimization. Subsequently, we identify a common family of acquisition functions, including EI and UCB, whose characteristics not only facilitate but justify use of greedy approaches for their maximization.

MetaReg: Towards Domain Generalization using Meta-Regularization

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #157

Yogesh Balaji · Swami Sankaranarayanan · Rama Chellappa

Training models that generalize to new domains at test time is a problem of fundamental importance in machine learning. In this work, we encode this notion of domain generalization using a novel regularization function. We pose the problem of finding such a regularization function in a Learning to Learn (or) meta-learning framework. The objective of domain generalization is explicitly modeled by learning a regularizer that makes the model trained on one domain to perform well on another domain. Experimental validations on computer vision and natural language datasets indicate that our method can learn regularizers that achieve good cross-domain generalization.

Transfer Learning with Neural AutoML

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #158

Catherine Wong · Neil Houlsby · Yifeng Lu · Andrea Gesmundo

We reduce the computational cost of Neural AutoML with transfer learning. AutoML relieves human effort by automating the design of ML algorithms. Neural AutoML has become popular for the design of deep learning architectures, however, this method has a high computation cost. To address this we propose Transfer Neural AutoML that uses knowledge from prior tasks to speed up network design. We extend RL-based architecture search methods to support parallel training on multiple tasks and then transfer the search strategy to new tasks. On language and image classification data, Transfer Neural AutoML reduces convergence time over single-task training by over an order of magnitude on many tasks.

Hierarchical Reinforcement Learning for Zero-shot Generalization with Subtask Dependencies

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #159

Sungryull Sohn · Junhyuk Oh · Honglak Lee

We introduce a new RL problem where the agent is required to generalize to a previously-unseen environment characterized by a subtask graph which describes a set of subtasks and their dependencies. Unlike existing hierarchical multitask RL approaches that explicitly describe what the agent should do at a high level, our problem only describes properties of subtasks and relationships among them, which requires the agent to perform complex reasoning to find the optimal subtask to execute. To solve this problem, we propose a neural subtask graph solver (NSGS) which encodes the subtask graph using a recursive neural network embedding. To overcome the difficulty of training, we propose a novel non-parametric gradient-based policy, graph reward propagation, to pre-train our NSGS agent and further finetune it through actor-critic method. The experimental results on two 2D visual domains show that our agent can perform complex reasoning to find a near-optimal way of executing the subtask graph and generalize well to the unseen subtask graphs. In addition, we compare our agent with a Monte-Carlo tree search (MCTS) method showing that our method is much more efficient than MCTS, and the performance of NSGS can be further improved by combining it with MCTS.

Lifelong Inverse Reinforcement Learning

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #160

Jorge A Mendez · Shashank Shivkumar · Eric Eaton

Methods for learning from demonstration (LfD) have shown success in acquiring behavior policies by imitating a user. However, even for a single task, LfD may require numerous demonstrations. For versatile agents that must learn many tasks via demonstration, this process would substantially burden the user if each task were learned in isolation. To address this challenge, we introduce the novel problem of lifelong learning from demonstration, which allows the agent to continually build upon knowledge learned from previously demonstrated tasks to accelerate the learning of new tasks, reducing the amount of demonstrations required. As one solution to this problem, we propose the first lifelong learning approach to inverse reinforcement learning, which learns consecutive tasks via demonstration, continually transferring knowledge between tasks to improve performance.

Safe Active Learning for Time-Series Modeling with Gaussian Processes

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #161

Christoph Zimmer · Mona Meister · Duy Nguyen-Tuong

Learning time-series models is useful for many applications, such as simulation and forecasting. In this study, we consider the problem of actively learning time-series models while taking given safety constraints into account. For time-series modeling we employ a Gaussian process with a nonlinear exogenous input structure. The proposed approach generates data appropriate for time series model learning, i.e. input and output trajectories, by dynamically exploring the input space. The approach parametrizes the input trajectory as consecutive trajectory sections, which are determined stepwise given safety requirements and past observations. We analyze the proposed algorithm and evaluate it empirically on a technical application. The results show the effectiveness of our approach in a realistic technical use case.

Online Structure Learning for Feed-Forward and Recurrent Sum-Product Networks

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #162

Agastya Kalra · Abdullah Rashwan · Wei-Shou Hsu · Pascal Poupart · Prashant Doshi · Georgios Trimponias

Sum-product networks have recently emerged as an attractive representation due to their dual view as a special type of deep neural network with clear semantics and a special type of probabilistic graphical model for which inference is always tractable. Those properties follow from some conditions (i.e., completeness and decomposability) that must be respected by the structure of the network. As a result, it is not easy to specify a valid sum-product network by hand and therefore structure learning techniques are typically used in practice. This paper describes a new online structure learning technique for feed-forward and recurrent SPNs. The algorithm is demonstrated on real-world datasets with continuous features for which it is not clear what network architecture might be best, including sequence datasets of varying length.

Preference Based Adaptation for Learning Objectives

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #163

Yao-Xiang Ding · Zhi-Hua Zhou

In many real-world learning tasks, it is hard to directly optimize the true performance measures, meanwhile choosing the right surrogate objectives is also difficult. Under this situation, it is desirable to incorporate an optimization of objective process into the learning loop based on weak modeling of the relationship between the true measure and the objective. In this work, we discuss the task of objective adaptation, in which the learner iteratively adapts the learning objective to the underlying true objective based on the preference feedback from an oracle. We show that when the objective can be linearly parameterized, this preference based learning problem can be solved by utilizing the dueling bandit model. A novel sampling based algorithm DL^2M is proposed to learn the optimal parameter, which enjoys strong theoretical guarantees and efficient empirical performance. To avoid learning a hypothesis from scratch after each objective function update, a boosting based hypothesis adaptation approach is proposed to efficiently adapt any pre-learned element hypothesis to the current objective. We apply the overall approach to multi-label learning, and show that the proposed approach achieves significant performance under various multi-label performance measures.

Byzantine Stochastic Gradient Descent

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #164

Dan Alistarh · Zeyuan Allen-Zhu · Jerry Li

This paper studies the problem of distributed stochastic optimization in an adversarial setting where, out of m machines which allegedly compute stochastic gradients every iteration, an α-fraction are Byzantine, and may behave adversarially. Our main result is a variant of stochastic gradient descent (SGD) which finds ε-approximate minimizers of convex functions in $T = \tilde{O}\big(\frac{1}{\varepsilon^2 m} + \frac{\alpha^2}{\varepsilon^2} \big)$ iterations. In contrast, traditional mini-batch SGD needs $T = O\big(\frac{1}{\varepsilon^2 m} \big)$ iterations, but cannot tolerate Byzantine failures. Further, we provide a lower bound showing that, up to logarithmic factors, our algorithm is information-theoretically optimal both in terms of sample complexity and time complexity.

Contextual bandits with surrogate losses: Margin bounds and efficient algorithms

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #165

Dylan Foster · Akshay Krishnamurthy

We use surrogate losses to obtain several new regret bounds and new algorithms for contextual bandit learning. Using the ramp loss, we derive a new margin-based regret bound in terms of standard sequential complexity measures of a benchmark class of real-valued regression functions. Using the hinge loss, we derive an efficient algorithm with a $\sqrt{dT}$-type mistake bound against benchmark policies induced by d-dimensional regressors. Under realizability assumptions, our results also yield classical regret bounds.

Online Learning of Quantum States

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #166

Scott Aaronson · Xinyi Chen · Elad Hazan · Satyen Kale · Ashwin Nayak

Suppose we have many copies of an unknown n-qubit state ρ. We measure some copies of ρ using a known two-outcome measurement E_1, then other copies using a measurement E_2, and so on. At each stage t, we generate a current hypothesis ω_t about the state ρ, using the outcomes of the previous measurements. We show that it is possible to do this in a way that guarantees that $|\trace(E_i \omega_t) - \trace(E_i\rho)|$, the error in our prediction for the next measurement, is at least eps at most $O(n / eps^2) $\ times. Even in the non-realizable setting---where there could be arbitrary noise in the measurement outcomes---we show how to output hypothesis states that incur at most $O(\sqrt {Tn}) $ excess loss over the best possible state on the first T measurements. These results generalize a 2007 theorem by Aaronson on the PAC-learnability of quantum states, to the online and regret-minimization settings. We give three different ways to prove our results---using convex optimization, quantum postselection, and sequential fat-shattering dimension---which have different advantages in terms of parameters and portability.

Horizon-Independent Minimax Linear Regression

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #167

Alan Malek · Peter Bartlett

We consider online linear regression: at each round, an adversary reveals a covariate vector, the learner predicts a real value, the adversary reveals a label, and the learner suffers the squared prediction error. The aim is to minimize the difference between the cumulative loss and that of the linear predictor that is best in hindsight. Previous work demonstrated that the minimax optimal strategy is easy to compute recursively from the end of the game; this requires the entire sequence of covariate vectors in advance. We show that, once provided with a measure of the scale of the problem, we can invert the recursion and play the minimax strategy without knowing the future covariates. Further, we show that this forward recursion remains optimal even against adaptively chosen labels and covariates, provided that the adversary adheres to a set of constraints that prevent misrepresentation of the scale of the problem. This strategy is horizon-independent in that the regret and minimax strategies depend on the size of the constraint set and not on the time-horizon, and hence it incurs no more regret than the optimal strategy that knows in advance the number of rounds of the game. We also provide an interpretation of the minimax algorithm as a follow-the-regularized-leader strategy with a data-dependent regularizer and obtain an explicit expression for the minimax regret.

Factored Bandits

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #168

Julian Zimmert · Yevgeny Seldin

We introduce the factored bandits model, which is a framework for learning with limited (bandit) feedback, where actions can be decomposed into a Cartesian product of atomic actions. Factored bandits incorporate rank-1 bandits as a special case, but significantly relax the assumptions on the form of the reward function. We provide an anytime algorithm for stochastic factored bandits and up to constants matching upper and lower regret bounds for the problem. Furthermore, we show that with a slight modification the proposed algorithm can be applied to utility based dueling bandits. We obtain an improvement in the additive terms of the regret bound compared to state of the art algorithms (the additive terms are dominating up to time horizons which are exponential in the number of arms).

A Model for Learned Bloom Filters and Optimizing by Sandwiching

Poster | Thu Dec 6th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #169

Michael Mitzenmacher

Recent work has suggested enhancing Bloom filters by using a pre-filter, based on applying machine learning to determine a function that models the data set the Bloom filter is meant to represent. Here we model such learned Bloom filters, with the following outcomes: (1) we clarify what guarantees can and cannot be associated with such a structure; (2) we show how to estimate what size the learning function must obtain in order to obtain improved performance; (3) we provide a simple method, sandwiching, for optimizing learned Bloom filters; and (4) we propose a design and analysis approach for a learned Bloomier filter, based on our modeling approach.

Designing Computer Systems for Software 2.0

Invited Talk | Thu Dec 6th 02:15 -- 03:05 PM @ Rooms 220 CDE

Kunle Olukotun

The use of machine learning to generate models from data is replacing traditional software development for many applications. This fundamental shift in how we develop software, known as Software 2.0, has provided dramatic improvements in the quality and ease of deployment for these applications. The continued success and expansion of the Software 2.0 approach must be powered by the availability of powerful, efficient and flexible computer systems that are tailored for machine learning applications. This talk will describe a design approach that optimizes computer systems to match the requirements of machine learning applications. The full-stack design approach integrates machine learning algorithms that are optimized for the characteristics of applications and the strengths of modern hardware, domain-specific languages and advanced compilation technology designed for programmability and performance, and hardware architectures that achieve both high flexibility and high energy efficiency.

Coffee Break

Break | Thu Dec 6th 03:05 -- 03:30 PM @

—

Robust Subspace Approximation in a Stream

Spotlight | Thu Dec 6th 03:30 -- 03:35 PM @ Room 220 CD

Roie Levin · Anish Prasad Sevekari · David Woodruff

We study robust subspace estimation in the streaming and distributed settings. Given a set of n data points {a_i}*{i=1}^n in R^d and an integer k, we wish to find a linear subspace S of dimension k for which sum_i M(dist(S, a_i)) is minimized, where dist(S,x) := min*{y in S} |x-y|_2, and M() is some loss function. When M is the identity function, S gives a subspace that is more robust to outliers than that provided by the truncated SVD. Though the problem is NP-hard, it is approximable within a (1+epsilon) factor in polynomial time when k and epsilon are constant. We give the first sublinear approximation algorithm for this problem in the turnstile streaming and arbitrary partition distributed models, achieving the same time guarantees as in the offline case. Our algorithm is the first based entirely on oblivious dimensionality reduction, and significantly simplifies prior methods for this problem, which held in neither the streaming nor distributed models.

Hyperbolic Neural Networks

Spotlight | Thu Dec 6th 03:30 -- 03:35 PM @ Room 220 E

Octavian Ganea · Gary Becigneul · Thomas Hofmann

Hyperbolic spaces have recently gained momentum in the context of machine learning due to their high capacity and tree-likeliness properties. However, the representational power of hyperbolic geometry is not yet on par with Euclidean geometry, firstly because of the absence of corresponding hyperbolic neural network layers. Here, we bridge this gap in a principled manner by combining the formalism of Möbius gyrovector spaces with the Riemannian geometry of the Poincaré model of hyperbolic spaces. As a result, we derive hyperbolic versions of important deep learning tools: multinomial logistic regression, feed-forward and recurrent neural networks. This allows to embed sequential data and perform classification in the hyperbolic space. Empirically, we show that, even if hyperbolic optimization tools are limited, hyperbolic sentence embeddings either outperform or are on par with their Euclidean variants on textual entailment and noisy-prefix recognition tasks.

A Simple Proximal Stochastic Gradient Method for Nonsmooth Nonconvex Optimization

Spotlight | Thu Dec 6th 03:30 -- 03:35 PM @ Room 517 CD

Zhize Li · Jian Li

We analyze stochastic gradient algorithms for optimizing nonconvex, nonsmooth finite-sum problems. In particular, the objective function is given by the summation of a differentiable (possibly nonconvex) component, together with a possibly non-differentiable but convex component. We propose a proximal stochastic gradient algorithm based on variance reduction, called ProxSVRG+. Our main contribution lies in the analysis of ProxSVRG+. It recovers several existing convergence results and improves/generalizes them (in terms of the number of stochastic gradient oracle calls and proximal oracle calls). In particular, ProxSVRG+ generalizes the best results given by the SCSG algorithm, recently proposed by [Lei et al., NIPS'17] for the smooth nonconvex case. ProxSVRG+ is also more straightforward than SCSG and yields simpler analysis. Moreover, ProxSVRG+ outperforms the deterministic proximal gradient descent (ProxGD) for a wide range of minibatch sizes, which partially solves an open problem proposed in [Reddi et al., NIPS'16]. Also, ProxSVRG+ uses much less proximal oracle calls than ProxSVRG [Reddi et al., NIPS'16]. Moreover, for nonconvex functions satisfied Polyak-\L{}ojasiewicz condition, we prove that ProxSVRG+ achieves a global linear convergence rate without restart unlike ProxSVRG. Thus, it can \emph{automatically} switch to the faster linear convergence in some regions as long as the objective function satisfies the PL condition locally in these regions. Finally, we conduct several experiments and the experimental results are consistent with the theoretical results.

Efficient nonmyopic batch active search

Spotlight | Thu Dec 6th 03:35 -- 03:40 PM @ Room 220 CD

Shali Jiang · Gustavo Malkomes · Matthew Abbott · Benjamin Moseley · Roman Garnett

Active search is a learning paradigm for actively identifying as many members of a given class as possible. A critical target scenario is high-throughput screening for scientific discovery, such as drug or materials discovery. In these settings, specialized instruments can often evaluate \emph{multiple} points simultaneously; however, all existing work on active search focuses on sequential acquisition. We bridge this gap, addressing batch active search from both the theoretical and practical perspective. We first derive the Bayesian optimal policy for this problem, then prove a lower bound on the performance gap between sequential and batch optimal policies: the ``cost of parallelization.'' We also propose novel, efficient batch policies inspired by state-of-the-art sequential policies, and develop an aggressive pruning technique that can dramatically speed up computation. We conduct thorough experiments on data from three application domains: a citation network, material science, and drug discovery, testing all proposed policies (14 total) with a wide range of batch sizes. Our results demonstrate that the empirical performance gap matches our theoretical bound, that nonmyopic policies usually significantly outperform myopic alternatives, and that diversity is an important consideration for batch policy design.

Norm matters: efficient and accurate normalization schemes in deep networks

Spotlight | Thu Dec 6th 03:35 -- 03:40 PM @ Room 220 E

Elad Hoffer · Ron Banner · Itay Golan · Daniel Soudry

Over the past few years, Batch-Normalization has been commonly used in deep networks, allowing faster training and high performance for a wide variety of applications. However, the reasons behind its merits remained unanswered, with several shortcomings that hindered its use for certain tasks. In this work, we present a novel view on the purpose and function of normalization methods and weight-decay, as tools to decouple weights' norm from the underlying optimized objective. This property highlights the connection between practices such as normalization, weight decay and learning-rate adjustments. We suggest several alternatives to the widely used \$L^2\$ batch-norm, using normalization in \$L^1\$ and \$L^\infty\$ spaces that can substantially improve numerical stability in low-precision implementations as well as provide computational and memory benefits. We demonstrate that such methods enable the first batch-norm alternative to work for half-precision implementations. Finally, we suggest a modification to weight-normalization, which improves its performance on large-scale tasks.

Stochastic Chebyshev Gradient Descent for Spectral Optimization

Spotlight | Thu Dec 6th 03:35 -- 03:40 PM @ Room 517 CD

Insu Han · Haim Avron · Jinwoo Shin

A large class of machine learning techniques requires the solution of optimization problems involving spectral functions of parametric matrices, e.g. log-determinant and nuclear norm. Unfortunately, computing the gradient of a spectral function is generally of cubic complexity, as such gradient descent methods are rather expensive for optimizing objectives involving the spectral function. Thus, one naturally turns to stochastic gradient methods in hope that they will provide a way to reduce or altogether avoid the computation of full gradients. However, here a new challenge appears: there is no straightforward way to compute unbiased stochastic gradients for spectral functions. In this paper, we develop unbiased stochastic gradients for spectral-sums, an important subclass of spectral functions. Our unbiased stochastic gradients are based on combining randomized trace estimators with stochastic truncation of the Chebyshev expansions. A careful design of the truncation distribution allows us to offer distributions that are variance-optimal, which is crucial for fast and stable convergence of stochastic gradient methods. We further leverage our proposed stochastic gradients to devise stochastic methods for objective functions involving spectral-sums, and rigorously analyze their convergence rate. The utility of our methods is demonstrated in numerical experiments.

Interactive Structure Learning with Structural Query-by-Committee

Spotlight | Thu Dec 6th 03:40 -- 03:45 PM @ Room 220 CD

Christopher Tosh · Sanjoy Dasgupta

In this work, we introduce interactive structure learning, a framework that unifies many different interactive learning tasks. We present a generalization of the query-by-committee active learning algorithm for this setting, and we study its consistency and rate of convergence, both theoretically and empirically, with and without noise.

Constructing Fast Network through Deconstruction of

Convolution

Spotlight | Thu Dec 6th 03:40 -- 03:45 PM @ Room 220 E

Yunho Jeon · Junmo Kim

Convolutional neural networks have achieved great success in various vision tasks; however, they incur heavy resource costs. By using deeper and wider networks, network accuracy can be improved rapidly. However, in an environment with limited resources (e.g., mobile applications), heavy networks may not be usable. This study shows that naive convolution can be deconstructed into a shift operation and pointwise convolution. To cope with various convolutions, we propose a new shift operation called active shift layer (ASL) that formulates the amount of shift as a learnable function with shift parameters. This new layer can be optimized end-to-end through backpropagation and it can provide optimal shift values. Finally, we apply this layer to a light and fast network that surpasses existing state-of-the-art networks.

LAG: Lazily Aggregated Gradient for Communication-Efficient Distributed Learning

Spotlight | Thu Dec 6th 03:40 -- 03:45 PM @ Room 517 CD

Tianyi Chen · Georgios Giannakis · Tao Sun · Wotao Yin

This paper presents a new class of gradient methods for distributed machine learning that adaptively skip the gradient calculations to learn with reduced communication and computation. Simple rules are designed to detect slowly-varying gradients and, therefore, trigger the reuse of outdated gradients. The resultant gradient-based algorithms are termed Lazily Aggregated Gradient --- justifying our acronym LAG used henceforth. Theoretically, the merits of this contribution are: i) the convergence rate is the same as batch gradient descent in strongly-convex, convex, and nonconvex cases; and, ii) if the distributed datasets are heterogeneous (quantified by certain measurable constants), the communication rounds needed to achieve a targeted accuracy are reduced thanks to the adaptive reuse of lagged gradients. Numerical experiments on both synthetic and real data corroborate a significant communication reduction compared to alternatives.

Contour location via entropy reduction leveraging multiple information sources

Spotlight | Thu Dec 6th 03:45 -- 03:50 PM @ Room 220 CD

Alexandre Marques · Remi Lam · Karen Willcox

We introduce an algorithm to locate contours of functions that are expensive to evaluate. The problem of locating contours arises in many applications, including classification, constrained optimization, and performance analysis of mechanical and dynamical systems (reliability, probability of failure, stability, etc.). Our algorithm locates contours using information from multiple sources, which are available in the form of relatively inexpensive, biased, and possibly noisy approximations to the original function. Considering multiple information sources can lead to significant cost savings. We also introduce the concept of contour entropy, a formal measure of uncertainty about the location of the zero contour of a function approximated by a statistical surrogate model. Our algorithm locates contours efficiently by maximizing the reduction of contour entropy per unit cost.

A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks

Spotlight | Thu Dec 6th 03:45 -- 03:50 PM @ Room 220 E

Kimin Lee · Kibok Lee · Honglak Lee · Jinwoo Shin

Detecting test samples drawn sufficiently far away from the training distribution statistically or adversarially is a fundamental requirement for deploying a good classifier in many real-world machine learning applications. However, deep neural networks with the softmax classifier are known to produce highly overconfident posterior distributions even for such abnormal samples. In this paper, we propose a simple yet effective method for detecting any abnormal samples, which is applicable to any pre-trained softmax neural classifier. We obtain the class conditional Gaussian distributions with respect to (low- and upper-level) features of the deep models under Gaussian discriminant analysis, which result in a confidence score based on the Mahalanobis distance. While most prior methods have been evaluated for detecting either out-of-distribution or adversarial samples, but not both, the proposed method achieves the state-of-the-art performances for both cases in our experiments. Moreover, we found that our proposed method is more robust in harsh cases, e.g., when the training dataset has noisy labels or small number of samples. Finally, we show that the proposed method enjoys broader usage by applying it to class-incremental learning: whenever out-of-distribution samples are detected, our classification rule can incorporate new classes well without further training deep models.

Low-rank Interaction with Sparse Additive Effects Model for Large Data Frames

Spotlight | Thu Dec 6th 03:45 -- 03:50 PM @ Room 517 CD

Geneviève Robin · Hoi-To Wai · Julie Josse · Olga Klopp · Eric Moulines

Many applications of machine learning involve the analysis of large data frames -- matrices collecting heterogeneous measurements (binary, numerical, counts, etc.) across samples -- with missing values. Low-rank models, as studied by Udell et al. (2016), are popular in this framework for tasks such as visualization, clustering and missing value imputation. Yet, available methods with statistical guarantees and efficient optimization do not allow explicit modeling of main additive effects such as row and column, or covariate effects. In this paper, we introduce a low-rank interaction and sparse additive effects (LORIS) model which combines matrix regression on a dictionary and low-rank design, to estimate main effects and interactions simultaneously. We provide statistical guarantees in the form of upper bounds on the estimation error of both components. Then, we introduce a mixed coordinate gradient descent (MCGD) method which provably converges sub-linearly to an optimal solution and is computationally efficient for large scale data sets. We show on simulated and survey data that the method has a clear advantage over current practices.

Non-delusional Q-learning and value-iteration

Oral | Thu Dec 6th 03:50 -- 04:05 PM @ Room 220 CD

Tyler Lu · Dale Schuurmans · Craig Boutilier

We identify a fundamental source of error in Q-learning and other forms of dynamic programming with function approximation. Delusional bias arises when the approximation architecture limits the class of expressible greedy policies. Since standard Q-updates make globally uncoordinated action choices with respect to the expressible policy class, inconsistent or even conflicting Q-value estimates can result, leading to pathological behaviour such as over/under-estimation, instability and even divergence. To solve this problem, we introduce a new notion of policy consistency and define a local backup process that ensures global consistency through the use of information sets---sets that record constraints on policies consistent with backed-up Q-values. We prove that both the model-based and model-free algorithms using this backup remove delusional bias, yielding the first known algorithms that guarantee optimal results under general conditions. These algorithms furthermore only require polynomially many information sets (from a potentially exponential support). Finally, we suggest other practical heuristics for value-iteration and Q-learning that attempt to reduce delusional bias.

Discovery of Latent 3D Keypoints via End-to-end Geometric Reasoning

Oral | Thu Dec 6th 03:50 -- 04:05 PM @ Room 220 E

Supasorn Suwajanakorn · Noah Snavely · Jonathan Tompson · Mohammad Norouzi

This paper presents KeypointNet, an end-to-end geometric reasoning framework to learn an optimal set of category-specific keypoints, along with their detectors to predict 3D keypoints in a single 2D input image. We demonstrate this framework on 3D pose estimation task by proposing a differentiable pose objective that seeks the optimal set of keypoints for recovering the relative pose between two views of an object. Our network automatically discovers a consistent set of keypoints across viewpoints of a single object as well as across all object instances of a given object class. Importantly, we find that our end-to-end approach using no ground-truth keypoint annotations outperforms a fully supervised baseline using the same neural network architecture for the pose estimation task. The discovered 3D keypoints across the car, chair, and plane categories of ShapeNet are visualized at https://keypoints.github.io/

Optimal Algorithms for Non-Smooth Distributed Optimization in Networks

Oral | Thu Dec 6th 03:50 -- 04:05 PM @ Room 517 CD

Kevin Scaman · Francis Bach · Sebastien Bubeck · Laurent Massoulié · Yin Tat Lee

In this work, we consider the distributed optimization of non-smooth convex functions using a network of computing units. We investigate this problem under two regularity assumptions: (1) the Lipschitz continuity of the global objective function, and (2) the Lipschitz continuity of local individual functions. Under the local regularity assumption, we provide the first optimal first-order decentralized algorithm called multi-step primal-dual (MSPD) and its corresponding optimal convergence rate. A notable aspect of this result is that, for non-smooth functions, while the dominant term of the error is in $O(1/\sqrt{t})$, the structure of the communication network only impacts a second-order term in $O(1/t)$, where t is time. In other words, the error due to limits in communication resources decreases at a fast rate even in the case of non-strongly-convex objective functions. Under the global regularity assumption, we provide a simple yet efficient algorithm called distributed randomized smoothing (DRS) based on a local smoothing of the objective function, and show that DRS is within a $d^{1/4}$ multiplicative factor of the optimal convergence rate, where d is the underlying dimension.

Policy-Conditioned Uncertainty Sets for Robust Markov Decision Processes

Spotlight | Thu Dec 6th 04:05 -- 04:10 PM @ Room 220 CD

Andrea Tirinzoni · Marek Petrik · Xiangli Chen · Brian Ziebart

What policy should be employed in a Markov decision process with uncertain parameters? Robust optimization answer to this question is to use rectangular uncertainty sets, which independently reflect available knowledge about each state, and then obtains a decision policy that maximizes expected reward for the worst-case decision process parameters from these uncertainty sets. While this rectangularity is convenient computationally and leads to tractable solutions, it often produces policies that are too conservative in practice, and does not facilitate knowledge transfer between portions of the state space or across related decision processes. In this work, we propose non-rectangular uncertainty sets that bound marginal moments of state-action features defined over entire trajectories through a decision process. This enables generalization to different portions of the state space while retaining appropriate uncertainty of the decision process. We develop algorithms for solving the resulting robust decision problems, which reduce to finding an optimal policy for a mixture of decision processes, and demonstrate the benefits of our approach experimentally.

Learning Libraries of Subroutines for Neurally–Guided Bayesian Program Induction

Spotlight | Thu Dec 6th 04:05 -- 04:10 PM @ Room 220 E

Kevin Ellis · Lucas Morales · Mathias Sablé-Meyer · Armando Solar-Lezama · Josh Tenenbaum

Successful approaches to program induction require a hand-engineered domain-specific language (DSL), constraining the space of allowed programs and imparting prior knowledge of the domain. We contribute a program induction algorithm that learns a DSL while jointly training a neural network to efficiently search for programs in the learned DSL. We use our model to synthesize functions on lists, edit text, and solve symbolic regression problems, showing how the model learns a domain-specific library of program components for expressing solutions to problems in the domain.

Direct Runge-Kutta Discretization Achieves Acceleration

Spotlight | Thu Dec 6th 04:05 -- 04:10 PM @ Room 517 CD

Jingzhao Zhang · Aryan Mokhtari · Suvrit Sra · Ali Jadbabaie

We study gradient-based optimization methods obtained by directly discretizing a second-order ordinary differential equation (ODE) related to the continuous limit of Nesterov's accelerated gradient method. When the function is smooth enough, we show that acceleration can be achieved by a stable discretization of this ODE using standard Runge-Kutta integrators. Specifically, we prove that under Lipschitz-gradient, convexity and order-$(s+2)$ differentiability assumptions, the sequence of iterates generated by discretizing the proposed second-order ODE converges to the optimal solution at a rate of $\mathcal{O}({N^{-2\frac{s}{s+1}}})$, where s is the order of the Runge-Kutta numerical integrator. Furthermore, we introduce a new local flatness condition on the objective, under which rates even faster than $\mathcal{O}(N^{-2})$ can be achieved with low-order integrators and only gradient information. Notably, this flatness condition is satisfied by several standard loss functions used in machine learning. We provide numerical experiments that verify the theoretical rates predicted by our results.

Learning convex bounds for linear quadratic control policy synthesis

Spotlight | Thu Dec 6th 04:10 -- 04:15 PM @ Room 220 CD

Jack Umenberger · Thomas Schön

Learning to make decisions from observed data in dynamic environments remains a problem of fundamental importance in a numbers of fields, from artificial intelligence and robotics, to medicine and finance. This paper concerns the problem of learning control policies for unknown linear dynamical systems so as to maximize a quadratic reward function. We present a method to optimize the expected value of the reward over the posterior distribution of the unknown system parameters, given data. The algorithm involves sequential convex programing, and enjoys reliable local convergence and robust stability guarantees. Numerical simulations and stabilization of a real-world inverted pendulum are used to demonstrate the approach, with strong performance and robustness properties observed in both.

Learning Loop Invariants for Program Verification

Spotlight | Thu Dec 6th 04:10 -- 04:15 PM @ Room 220 E

Xujie Si · Hanjun Dai · Mukund Raghothaman · Mayur Naik · Le Song

A fundamental problem in program verification concerns inferring loop invariants. The problem is undecidable and even practical instances are challenging. Inspired by how human experts construct loop invariants, we propose a reasoning framework Code2Inv that constructs the solution by multi-step decision making and querying an external program graph memory block. By training with reinforcement learning, Code2Inv captures rich program features and avoids the need for ground truth solutions as supervision. Compared to previous learning tasks in domains with graph-structured data, it addresses unique challenges, such as a binary objective function and an extremely sparse reward that is given by an automated theorem prover only after the complete loop invariant is proposed. We evaluate Code2Inv on a suite of 133 benchmark problems and compare it to three state-of-the-art systems. It solves 106 problems compared to 73 by a stochastic search-based system, 77 by a heuristic search-based system, and 100 by a decision tree learning-based system. Moreover, the strategy learned can be generalized to new programs: compared to solving new instances from scratch, the pre-trained agent is more sample efficient in finding solutions.

Limited Memory Kelley's Method Converges for Composite Convex and Submodular Objectives

Spotlight | Thu Dec 6th 04:10 -- 04:15 PM @ Room 517 CD

Song Zhou · Swati Gupta · Madeleine Udell

The original simplicial method (OSM), a variant of the classic Kelley’s cutting plane method, has been shown to converge to the minimizer of a composite convex and submodular objective, though no rate of convergence for this method was known. Moreover, OSM is required to solve subproblems in each iteration whose size grows linearly in the number of iterations. We propose a limited memory version of Kelley’s method (L-KM) and of OSM that requires limited memory (at most n+ 1 constraints for an n-dimensional problem) independent of the iteration. We prove convergence for L-KM when the convex part of the objective g is strongly convex and show it converges linearly when g is also smooth. Our analysis relies on duality between minimization of the composite convex and submodular objective and minimization of a convex function over the submodular base polytope. We introduce a limited memory version, L-FCFW, of the Fully-Corrective Frank-Wolfe (FCFW) method with approximate correction, to solve the dual problem. We show that L-FCFW and L-KM are dual algorithms that produce the same sequence of iterates; hence both converge linearly (when g is smooth and strongly convex) and with limited memory. We propose L-KM to minimize composite convex and submodular objectives; however, our results on L-FCFW hold for general polytopes and may be of independent interest.

Multiple-Step Greedy Policies in Approximate and Online Reinforcement Learning

Spotlight | Thu Dec 6th 04:15 -- 04:20 PM @ Room 220 CD

Yonathan Efroni · Gal Dalal · Bruno Scherrer · Shie Mannor

Multiple-step lookahead policies have demonstrated high empirical competence in Reinforcement Learning, via the use of Monte Carlo Tree Search or Model Predictive Control. In a recent work (Efroni et al., 2018), multiple-step greedy policies and their use in vanilla Policy Iteration algorithms were proposed and analyzed. In this work, we study multiple-step greedy algorithms in more practical setups. We begin by highlighting a counter-intuitive difficulty, arising with soft-policy updates: even in the absence of approximations, and contrary to the 1-step-greedy case, monotonic policy improvement is not guaranteed unless the update stepsize is sufficiently large. Taking particular care about this difficulty, we formulate and analyze online and approximate algorithms that use such a multi-step greedy operator.

DeepProbLog: Neural Probabilistic Logic Programming

Spotlight | Thu Dec 6th 04:15 -- 04:20 PM @ Room 220 E

Robin Manhaeve · Sebastijan Dumancic · Angelika Kimmig · Thomas Demeester · Luc De Raedt

We introduce DeepProbLog, a probabilistic logic programming language that incorporates deep learning by means of neural predicates. We show how existing inference and learning techniques can be adapted for the new language. Our experiments demonstrate that DeepProbLog supports (i) both symbolic and subsymbolic representations and inference, (ii) program induction, (iii) probabilistic (logic) programming, and (iv) (deep) learning from examples. To the best of our knowledge, this work is the first to propose a framework where general-purpose neural networks and expressive probabilistic-logical modeling and reasoning are integrated in a way that exploits the full expressiveness and strengths of both worlds and can be trained end-to-end based on examples.

(Probably) Concave Graph Matching

Spotlight | Thu Dec 6th 04:15 -- 04:20 PM @ Room 517 CD

Haggai Maron · Yaron Lipman

In this paper we address the graph matching problem. Following the recent works of \cite{zaslavskiy2009path,Vestner2017} we analyze and generalize the idea of concave relaxations. We introduce the concepts of \emph{conditionally concave} and \emph{probably conditionally concave} energies on polytopes and show that they encapsulate many instances of the graph matching problem, including matching Euclidean graphs and graphs on surfaces. We further prove that local minima of probably conditionally concave energies on general matching polytopes (\eg, doubly stochastic) are with high probability extreme points of the matching polytope (\eg, permutations).

Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models

Spotlight | Thu Dec 6th 04:20 -- 04:25 PM @ Room 220 CD

Kurtland Chua · Roberto Calandra · Rowan McAllister · Sergey Levine

Model-based reinforcement learning (RL) algorithms can attain excellent sample efficiency, but often lag behind the best model-free algorithms in terms of asymptotic performance. This is especially true with high-capacity parametric function approximators, such as deep networks. In this paper, we study how to bridge this gap, by employing uncertainty-aware dynamics models. We propose a new algorithm called probabilistic ensembles with trajectory sampling (PETS) that combines uncertainty-aware deep network dynamics models with sampling-based uncertainty propagation. Our comparison to state-of-the-art model-based and model-free deep RL algorithms shows that our approach matches the asymptotic performance of model-free algorithms on several challenging benchmark tasks, while requiring significantly fewer samples (e.g. 8 and 125 times fewer samples than Soft Actor Critic and Proximal Policy Optimization respectively on the half-cheetah task).

Learning to Infer Graphics Programs from Hand-Drawn Images

Spotlight | Thu Dec 6th 04:20 -- 04:25 PM @ Room 220 E

Kevin Ellis · Daniel Ritchie · Armando Solar-Lezama · Josh Tenenbaum

We introduce a model that learns to convert simple hand drawings into graphics programs written in a subset of \LaTeX.~The model combines techniques from deep learning and program synthesis. We learn a convolutional neural network that proposes plausible drawing primitives that explain an image. These drawing primitives are a specification (spec) of what the graphics program needs to draw. We learn a model that uses program synthesis techniques to recover a graphics program from that spec. These programs have constructs like variable bindings, iterative loops, or simple kinds of conditionals. With a graphics program in hand, we can correct errors made by the deep network and extrapolate drawings.

Graph Oracle Models, Lower Bounds, and Gaps for Parallel Stochastic Optimization

Spotlight | Thu Dec 6th 04:20 -- 04:25 PM @ Room 517 CD

Blake Woodworth · Jialei Wang · Adam Smith · Brendan McMahan · Nati Srebro

We suggest a general oracle-based framework that captures parallel stochastic optimization in different parallelization settings described by a dependency graph, and derive generic lower bounds in terms of this graph. We then use the framework and derive lower bounds to study several specific parallel optimization settings, including delayed updates and parallel processing with intermittent communication. We highlight gaps between lower and upper bounds on the oracle complexity, and cases where the ``natural'' algorithms are not known to be optimal.

Sample-Efficient Reinforcement Learning with Stochastic Ensemble Value Expansion

Oral | Thu Dec 6th 04:25 -- 04:40 PM @ Room 220 CD

Jacob Buckman · Danijar Hafner · George Tucker · Eugene Brevdo · Honglak Lee

There is growing interest in combining model-free and model-based approaches in reinforcement learning with the goal of achieving the high performance of model-free algorithms with low sample complexity. This is difficult because an imperfect dynamics model can degrade the performance of the learning algorithm, and in sufficiently complex environments, the dynamics model will always be imperfect. As a result, a key challenge is to combine model-based approaches with model-free learning in such a way that errors in the model do not degrade performance. We propose stochastic ensemble value expansion (STEVE), a novel model-based technique that addresses this issue. By dynamically interpolating between model rollouts of various horizon lengths, STEVE ensures that the model is only utilized when doing so does not introduce significant errors. Our approach outperforms model-free baselines on challenging continuous control benchmarks with an order-of-magnitude increase in sample efficiency.

Learning to Reconstruct Shapes from Unseen Classes

Oral | Thu Dec 6th 04:25 -- 04:40 PM @ Room 220 E

Xiuming Zhang · Zhoutong Zhang · Chengkai Zhang · Josh Tenenbaum · Bill Freeman · Jiajun Wu

From a single image, humans are able to perceive the full 3D shape of an object by exploiting learned shape priors from everyday life. Contemporary single-image 3D reconstruction algorithms aim to solve this task in a similar fashion, but often end up with priors that are highly biased by training classes. Here we present an algorithm, Generalizable Reconstruction (GenRe), designed to capture more generic, class-agnostic shape priors. We achieve this with an inference network and training procedure that combine 2.5D representations of visible surfaces (depth and silhouette), spherical shape representations of both visible and non-visible surfaces, and 3D voxel-based representations, in a principled manner that exploits the causal structure of how 3D shapes give rise to 2D images. Experiments demonstrate that GenRe performs well on single-view shape reconstruction, and generalizes to diverse novel objects from categories not seen during training.

Smoothed analysis of the low-rank approach for smooth semidefinite programs

Oral | Thu Dec 6th 04:25 -- 04:40 PM @ Room 517 CD

Thomas Pumir · Samy Jelassi · Nicolas Boumal

We consider semidefinite programs (SDPs) of size n with equality constraints. In order to overcome scalability issues, Burer and Monteiro proposed a factorized approach based on optimizing over a matrix Y of size $n\times k$ such that $X=YY^*$ is the SDP variable. The advantages of such formulation are twofold: the dimension of the optimization variable is reduced, and positive semidefiniteness is naturally enforced. However, optimization in Y is non-convex. In prior work, it has been shown that, when the constraints on the factorized variable regularly define a smooth manifold, provided k is large enough, for almost all cost matrices, all second-order stationary points (SOSPs) are optimal. Importantly, in practice, one can only compute points which approximately satisfy necessary optimality conditions, leading to the question: are such points also approximately optimal? To this end, and under similar assumptions, we use smoothed analysis to show that approximate SOSPs for a randomly perturbed objective function are approximate global optima, with k scaling like the square root of the number of constraints (up to log factors). We particularize our results to an SDP relaxation of phase retrieval.

Bilevel learning of the Group Lasso structure

Spotlight | Thu Dec 6th 04:40 -- 04:45 PM @ Room 220 CD

Jordan Frecon · Saverio Salzo · Massimiliano Pontil

Regression with group-sparsity penalty plays a central role in high-dimensional prediction problems. Most of existing methods require the group structure to be known a priori. In practice, this may be a too strong assumption, potentially hampering the effectiveness of the regularization method. To circumvent this issue, we present a method to estimate the group structure by means of a continuous bilevel optimization problem where the data is split into training and validation sets. Our approach relies on an approximation scheme where the lower level problem is replaced by a smooth dual forward-backward algorithm with Bregman distances. We provide guarantees regarding the convergence of the approximate procedure to the exact problem and demonstrate the well behaviour of the proposed method on synthetic experiments. Finally, a preliminary application to genes expression data is tackled with the purpose of unveiling functional groups.

Improving Neural Program Synthesis with Inferred Execution Traces

Spotlight | Thu Dec 6th 04:40 -- 04:45 PM @ Room 220 E

Richard Shin · Illia Polosukhin · Dawn Song

The task of program synthesis, or automatically generating programs that are consistent with a provided specification, remains a challenging task in artificial intelligence. As in other fields of AI, deep learning-based end-to-end approaches have made great advances in program synthesis. However, more so than other fields such as computer vision, program synthesis provides greater opportunities to explicitly exploit structured information such as execution traces, which contain a superset of the information input/output pairs. While they are highly useful for program synthesis, as execution traces are more difficult to obtain than input/output pairs, we use the insight that we can split the process into two parts: infer the trace from the input/output example, then infer the program from the trace. This simple modification leads to state-of-the-art results in program synthesis in the Karel domain, improving accuracy to 81.3% from the 77.12% of prior work.

Wasserstein Distributionally Robust Kalman Filtering

Spotlight | Thu Dec 6th 04:40 -- 04:45 PM @ Room 517 CD

Soroosh Shafieezadeh Abadeh · Viet Anh Nguyen · Daniel Kuhn · Peyman Mohajerin Esfahani

We study a distributionally robust mean square error estimation problem over a nonconvex Wasserstein ambiguity set containing only normal distributions. We show that the optimal estimator and the least favorable distribution form a Nash equilibrium. Despite the non-convex nature of the ambiguity set, we prove that the estimation problem is equivalent to a tractable convex program. We further devise a Frank-Wolfe algorithm for this convex program whose direction-searching subproblem can be solved in a quasi-closed form. Using these ingredients, we introduce a distributionally robust Kalman filter that hedges against model risk.

Binary Classification from Positive-Confidence Data

Spotlight | Thu Dec 6th 04:45 -- 04:50 PM @ Room 220 CD

Takashi Ishida · Gang Niu · Masashi Sugiyama

Can we learn a binary classifier from only positive data, without any negative data or unlabeled data? We show that if one can equip positive data with confidence (positive-confidence), one can successfully learn a binary classifier, which we name positive-confidence (Pconf) classification. Our work is related to one-class classification which is aimed at "describing" the positive class by clustering-related methods, but one-class classification does not have the ability to tune hyper-parameters and their aim is not on "discriminating" positive and negative classes. For the Pconf classification problem, we provide a simple empirical risk minimization framework that is model-independent and optimization-independent. We theoretically establish the consistency and an estimation error bound, and demonstrate the usefulness of the proposed method for training deep neural networks through experiments.

ResNet with one-neuron hidden layers is a Universal Approximator

Spotlight | Thu Dec 6th 04:45 -- 04:50 PM @ Room 220 E

Hongzhou Lin · Stefanie Jegelka

We demonstrate that a very deep ResNet with stacked modules that have one neuron per hidden layer and ReLU activation functions can uniformly approximate any Lebesgue integrable function in d dimensions, i.e. \ell_1(R^d). Due to the identity mapping inherent to ResNets, our network has alternating layers of dimension one and d. This stands in sharp contrast to fully connected networks, which are not universal approximators if their width is the input dimension d [21,11]. Hence, our result implies an increase in representational power for narrow deep networks by the ResNet architecture.

Decentralize and Randomize: Faster Algorithm for Wasserstein Barycenters

Spotlight | Thu Dec 6th 04:45 -- 04:50 PM @ Room 517 CD

Pavel Dvurechenskii · Darina Dvinskikh · Alexander Gasnikov · Cesar Uribe · Angelia Nedich

We study the decentralized distributed computation of discrete approximations for the regularized Wasserstein barycenter of a finite set of continuous probability measures distributedly stored over a network. We assume there is a network of agents/machines/computers, and each agent holds a private continuous probability measure and seeks to compute the barycenter of all the measures in the network by getting samples from its local measure and exchanging information with its neighbors. Motivated by this problem, we develop, and analyze, a novel accelerated primal-dual stochastic gradient method for general stochastic convex optimization problems with linear equality constraints. Then, we apply this method to the decen- tralized distributed optimization setting to obtain a new algorithm for the distributed semi-discrete regularized Wasserstein barycenter problem. Moreover, we show explicit non-asymptotic complexity for the proposed algorithm. Finally, we show the effectiveness of our method on the distributed computation of the regularized Wasserstein barycenter of univariate Gaussian and von Mises distributions, as well as some applications to image aggregation.

Fully Understanding The Hashing Trick

Spotlight | Thu Dec 6th 04:50 -- 04:55 PM @ Room 220 CD

Lior Kamma · Casper B. Freksen · Kasper Green Larsen

Feature hashing, also known as {\em the hashing trick}, introduced by Weinberger et al. (2009), is one of the key techniques used in scaling-up machine learning algorithms. Loosely speaking, feature hashing uses a random sparse projection matrix $A : \mathbb{R}^n \to \mathbb{R}^m$ (where $m \ll n$) in order to reduce the dimension of the data from n to m while approximately preserving the Euclidean norm. Every column of A contains exactly one non-zero entry, equals to either -1 or 1.

Weinberger et al. showed tail bounds on $|Ax|_2^2$. Specifically they showed that for every ε, δ, if $|x|_{\infty} / |x|_2$ is sufficiently small, and m is sufficiently large, then \begin{equation}\Pr[\; | \;|Ax|_2^2 - |x|_2^2\; | < \varepsilon |x|_2^2 \;] \ge 1 - \delta \;.\end{equation} These bounds were later extended by Dasgupta et al. (2010) and most recently refined by Dahlgaard et al. (2017), however, the true nature of the performance of this key technique, and specifically the correct tradeoff between the pivotal parameters $|x|_{\infty} / |x|_2, m, \varepsilon, \delta$ remained an open question.

We settle this question by giving tight asymptotic bounds on the exact tradeoff between the central parameters, thus providing a complete understanding of the performance of feature hashing. We complement the asymptotic bound with empirical data, which shows that the constants "hiding" in the asymptotic notation are, in fact, very close to 1, thus further illustrating the tightness of the presented bounds in practice.

Towards Understanding Learning Representations: To What Extent Do Different Neural Networks Learn the Same Representation

Spotlight | Thu Dec 6th 04:50 -- 04:55 PM @ Room 220 E

Liwei Wang · Lunjia Hu · Jiayuan Gu · Zhiqiang Hu · Yue Wu · Kun He · John Hopcroft

It is widely believed that learning good representations is one of the main reasons for the success of deep neural networks. Although highly intuitive, there is a lack of theory and systematic approach quantitatively characterizing what representations do deep neural networks learn. In this work, we move a tiny step towards a theory and better understanding of the representations. Specifically, we study a simpler problem: How similar are the representations learned by two networks with identical architecture but trained from different initializations. We develop a rigorous theory based on the neuron activation subspace match model. The theory gives a complete characterization of the structure of neuron activation subspace matches, where the core concepts are maximum match and simple match which describe the overall and the finest similarity between sets of neurons in two networks respectively. We also propose efficient algorithms to find the maximum match and simple matches. Finally, we conduct extensive experiments using our algorithms. Experimental results suggest that, surprisingly, representations learned by the same convolutional layers of networks trained from different initializations are not as similar as prevalently expected, at least in terms of subspace match.

Robust Hypothesis Testing Using Wasserstein Uncertainty Sets

Spotlight | Thu Dec 6th 04:50 -- 04:55 PM @ Room 517 CD

RUI GAO · Liyan Xie · Yao Xie · Huan Xu

We develop a novel computationally efficient and general framework for robust hypothesis testing. The new framework features a new way to construct uncertainty sets under the null and the alternative distributions, which are sets centered around the empirical distribution defined via Wasserstein metric, thus our approach is data-driven and free of distributional assumptions. We develop a convex safe approximation of the minimax formulation and show that such approximation renders a nearly-optimal detector among the family of all possible tests. By exploiting the structure of the least favorable distribution, we also develop a tractable reformulation of such approximation, with complexity independent of the dimension of observation space and can be nearly sample-size-independent in general. Real-data example using human activity data demonstrated the excellent performance of the new robust detector.

Support Recovery for Orthogonal Matching Pursuit: Upper and Lower bounds

Spotlight | Thu Dec 6th 04:55 -- 05:00 PM @ Room 220 CD

Raghav Somani · Chirag Gupta · Prateek Jain · Praneeth Netrapalli

This paper studies the problem of sparse regression where the goal is to learn a sparse vector that best optimizes a given objective function. Under the assumption that the objective function satisfies restricted strong convexity (RSC), we analyze orthogonal matching pursuit (OMP), a greedy algorithm that is used heavily in applications, and obtain support recovery result as well as a tight generalization error bound for OMP. Furthermore, we obtain lower bounds for OMP, showing that both our results on support recovery and generalization error are tight up to logarithmic factors. To the best of our knowledge, these support recovery and generalization bounds are the first such matching upper and lower bounds (up to logarithmic factors) for {\em any} sparse regression algorithm under the RSC assumption.

Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels

Spotlight | Thu Dec 6th 04:55 -- 05:00 PM @ Room 220 E

Zhilu Zhang · Mert Sabuncu

Deep neural networks (DNNs) have achieved tremendous success in a variety of applications across many disciplines. Yet, their superior performance comes with the expensive cost of requiring correctly annotated large-scale datasets. Moreover, due to DNNs' rich capacity, errors in training labels can hamper performance. To combat this problem, mean absolute error (MAE) has recently been proposed as a noise-robust alternative to the commonly-used categorical cross entropy (CCE) loss. However, as we show in this paper, MAE can perform poorly with DNNs and large-scale datasets. Here, we present a theoretically grounded set of noise-robust loss functions that can be seen as a generalization of MAE and CCE. Proposed loss functions can be readily applied with any existing DNN architecture and algorithm, while yielding good performance in a wide range of noisy label scenarios. We report results from experiments conducted with CIFAR-10, CIFAR-100 and FASHION-MNIST datasets and synthetically generated noisy labels.

Convergence of Cubic Regularization for Nonconvex Optimization under KL Property

Spotlight | Thu Dec 6th 04:55 -- 05:00 PM @ Room 517 CD

Yi Zhou · Zhe Wang · Yingbin Liang

Cubic-regularized Newton's method (CR) is a popular algorithm that guarantees to produce a second-order stationary solution for solving nonconvex optimization problems. However, existing understandings of convergence rate of CR are conditioned on special types of geometrical properties of the objective function. In this paper, we explore the asymptotic convergence rate of CR by exploiting the ubiquitous Kurdyka-Lojasiewicz (KL) property of the nonconvex objective functions. In specific, we characterize the asymptotic convergence rate of various types of optimality measures for CR including function value gap, variable distance gap, gradient norm and least eigenvalue of the Hessian matrix. Our results fully characterize the diverse convergence behaviors of these optimality measures in the full parameter regime of the KL property. Moreover, we show that the obtained asymptotic convergence rates of CR are order-wise faster than those of first-order gradient descent algorithms under the KL property.

Streamlining Variational Inference for Constraint Satisfaction Problems

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #1

Aditya Grover · Tudor Achim · Stefano Ermon

Several algorithms for solving constraint satisfaction problems are based on survey propagation, a variational inference scheme used to obtain approximate marginal probability estimates for variable assignments. These marginals correspond to how frequently each variable is set to true among satisfying assignments, and are used to inform branching decisions during search; however, marginal estimates obtained via survey propagation are approximate and can be self-contradictory. We introduce a more general branching strategy based on streamlining constraints, which sidestep hard assignments to variables. We show that streamlined solvers consistently outperform decimation-based solvers on random k-SAT instances for several problem sizes, shrinking the gap between empirical performance and theoretical limits of satisfiability by 16.3% on average for k = 3, 4, 5, 6.

Robust Hypothesis Testing Using Wasserstein Uncertainty Sets

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #2

RUI GAO · Liyan Xie · Yao Xie · Huan Xu

We develop a novel computationally efficient and general framework for robust hypothesis testing. The new framework features a new way to construct uncertainty sets under the null and the alternative distributions, which are sets centered around the empirical distribution defined via Wasserstein metric, thus our approach is data-driven and free of distributional assumptions. We develop a convex safe approximation of the minimax formulation and show that such approximation renders a nearly-optimal detector among the family of all possible tests. By exploiting the structure of the least favorable distribution, we also develop a tractable reformulation of such approximation, with complexity independent of the dimension of observation space and can be nearly sample-size-independent in general. Real-data example using human activity data demonstrated the excellent performance of the new robust detector.

Smoothed analysis of the low-rank approach for smooth semidefinite programs

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #3

Thomas Pumir · Samy Jelassi · Nicolas Boumal

We consider semidefinite programs (SDPs) of size \$n\$ with equality constraints. In order to overcome scalability issues, Burer and Monteiro proposed a factorized approach based on optimizing over a matrix \$Y\$ of size \$n\times k\$ such that \$X=YY^*\$ is the SDP variable. The advantages of such formulation are twofold: the dimension of the optimization variable is reduced, and positive semidefiniteness is naturally enforced. However, optimization in \$Y\$ is non-convex. In prior work, it has been shown that, when the constraints on the factorized variable regularly define a smooth manifold, provided \$k\$ is large enough, for almost all cost matrices, all second-order stationary points (SOSPs) are optimal. Importantly, in practice, one can only compute points which approximately satisfy necessary optimality conditions, leading to the question: are such points also approximately optimal? To this end, and under similar assumptions, we use smoothed analysis to show that approximate SOSPs for a randomly perturbed objective function are approximate global optima, with \$k\$ scaling like the square root of the number of constraints (up to log factors). We particularize our results to an SDP relaxation of phase retrieval.

Convergence of Cubic Regularization for Nonconvex Optimization under KL Property

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #4

Yi Zhou · Zhe Wang · Yingbin Liang

Cubic-regularized Newton's method (CR) is a popular algorithm that guarantees to produce a second-order stationary solution for solving nonconvex optimization problems. However, existing understandings of convergence rate of CR are conditioned on special types of geometrical properties of the objective function. In this paper, we explore the asymptotic convergence rate of CR by exploiting the ubiquitous Kurdyka-Lojasiewicz (KL) property of the nonconvex objective functions. In specific, we characterize the asymptotic convergence rate of various types of optimality measures for CR including function value gap, variable distance gap, gradient norm and least eigenvalue of the Hessian matrix. Our results fully characterize the diverse convergence behaviors of these optimality measures in the full parameter regime of the KL property. Moreover, we show that the obtained asymptotic convergence rates of CR are order-wise faster than those of first-order gradient descent algorithms under the KL property.

A Simple Proximal Stochastic Gradient Method for Nonsmooth Nonconvex Optimization

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #5

Zhize Li · Jian Li

We analyze stochastic gradient algorithms for optimizing nonconvex, nonsmooth finite-sum problems. In particular, the objective function is given by the summation of a differentiable (possibly nonconvex) component, together with a possibly non-differentiable but convex component. We propose a proximal stochastic gradient algorithm based on variance reduction, called ProxSVRG+. Our main contribution lies in the analysis of ProxSVRG+. It recovers several existing convergence results and improves/generalizes them (in terms of the number of stochastic gradient oracle calls and proximal oracle calls). In particular, ProxSVRG+ generalizes the best results given by the SCSG algorithm, recently proposed by [Lei et al., NIPS'17] for the smooth nonconvex case. ProxSVRG+ is also more straightforward than SCSG and yields simpler analysis. Moreover, ProxSVRG+ outperforms the deterministic proximal gradient descent (ProxGD) for a wide range of minibatch sizes, which partially solves an open problem proposed in [Reddi et al., NIPS'16]. Also, ProxSVRG+ uses much less proximal oracle calls than ProxSVRG [Reddi et al., NIPS'16]. Moreover, for nonconvex functions satisfied Polyak-\L{}ojasiewicz condition, we prove that ProxSVRG+ achieves a global linear convergence rate without restart unlike ProxSVRG. Thus, it can \emph{automatically} switch to the faster linear convergence in some regions as long as the objective function satisfies the PL condition locally in these regions. Finally, we conduct several experiments and the experimental results are consistent with the theoretical results.

Stochastic Chebyshev Gradient Descent for Spectral Optimization

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #6

Insu Han · Haim Avron · Jinwoo Shin

A large class of machine learning techniques requires the solution of optimization problems involving spectral functions of parametric matrices, e.g. log-determinant and nuclear norm. Unfortunately, computing the gradient of a spectral function is generally of cubic complexity, as such gradient descent methods are rather expensive for optimizing objectives involving the spectral function. Thus, one naturally turns to stochastic gradient methods in hope that they will provide a way to reduce or altogether avoid the computation of full gradients. However, here a new challenge appears: there is no straightforward way to compute unbiased stochastic gradients for spectral functions. In this paper, we develop unbiased stochastic gradients for spectral-sums, an important subclass of spectral functions. Our unbiased stochastic gradients are based on combining randomized trace estimators with stochastic truncation of the Chebyshev expansions. A careful design of the truncation distribution allows us to offer distributions that are variance-optimal, which is crucial for fast and stable convergence of stochastic gradient methods. We further leverage our proposed stochastic gradients to devise stochastic methods for objective functions involving spectral-sums, and rigorously analyze their convergence rate. The utility of our methods is demonstrated in numerical experiments.

Proximal SCOPE for Distributed Sparse Learning

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #7

Shenyi Zhao · Gong-Duo Zhang · Ming-Wei Li · Wu-Jun Li

Distributed sparse learning with a cluster of multiple machines has attracted much attention in machine learning, especially for large-scale applications with high-dimensional data. One popular way to implement sparse learning is to use L1 regularization. In this paper, we propose a novel method, called proximal SCOPE (pSCOPE), for distributed sparse learning with L1 regularization. pSCOPE is based on a cooperative autonomous local learning (CALL) framework. In the CALL framework of pSCOPE, we find that the data partition affects the convergence of the learning procedure, and subsequently we define a metric to measure the goodness of a data partition. Based on the defined metric, we theoretically prove that pSCOPE is convergent with a linear convergence rate if the data partition is good enough. We also prove that better data partition implies faster convergence rate. Furthermore, pSCOPE is also communication efficient. Experimental results on real data sets show that pSCOPE can outperform other state-of-the-art distributed methods for sparse learning.

LAG: Lazily Aggregated Gradient for Communication-Efficient Distributed Learning

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #8

Tianyi Chen · Georgios Giannakis · Tao Sun · Wotao Yin

This paper presents a new class of gradient methods for distributed machine learning that adaptively skip the gradient calculations to learn with reduced communication and computation. Simple rules are designed to detect slowly-varying gradients and, therefore, trigger the reuse of outdated gradients. The resultant gradient-based algorithms are termed Lazily Aggregated Gradient --- justifying our acronym LAG used henceforth. Theoretically, the merits of this contribution are: i) the convergence rate is the same as batch gradient descent in strongly-convex, convex, and nonconvex cases; and, ii) if the distributed datasets are heterogeneous (quantified by certain measurable constants), the communication rounds needed to achieve a targeted accuracy are reduced thanks to the adaptive reuse of lagged gradients. Numerical experiments on both synthetic and real data corroborate a significant communication reduction compared to alternatives.

Direct Runge-Kutta Discretization Achieves Acceleration

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #9

Jingzhao Zhang · Aryan Mokhtari · Suvrit Sra · Ali Jadbabaie

We study gradient-based optimization methods obtained by directly discretizing a second-order ordinary differential equation (ODE) related to the continuous limit of Nesterov's accelerated gradient method. When the function is smooth enough, we show that acceleration can be achieved by a stable discretization of this ODE using standard Runge-Kutta integrators. Specifically, we prove that under Lipschitz-gradient, convexity and order-$(s+2)$ differentiability assumptions, the sequence of iterates generated by discretizing the proposed second-order ODE converges to the optimal solution at a rate of $\mathcal{O}({N^{-2\frac{s}{s+1}}})$, where s is the order of the Runge-Kutta numerical integrator. Furthermore, we introduce a new local flatness condition on the objective, under which rates even faster than $\mathcal{O}(N^{-2})$ can be achieved with low-order integrators and only gradient information. Notably, this flatness condition is satisfied by several standard loss functions used in machine learning. We provide numerical experiments that verify the theoretical rates predicted by our results.

Optimal Algorithms for Non-Smooth Distributed Optimization in Networks

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #10

Kevin Scaman · Francis Bach · Sebastien Bubeck · Laurent Massoulié · Yin Tat Lee

In this work, we consider the distributed optimization of non-smooth convex functions using a network of computing units. We investigate this problem under two regularity assumptions: (1) the Lipschitz continuity of the global objective function, and (2) the Lipschitz continuity of local individual functions. Under the local regularity assumption, we provide the first optimal first-order decentralized algorithm called multi-step primal-dual (MSPD) and its corresponding optimal convergence rate. A notable aspect of this result is that, for non-smooth functions, while the dominant term of the error is in $O(1/\sqrt{t})$, the structure of the communication network only impacts a second-order term in $O(1/t)$, where t is time. In other words, the error due to limits in communication resources decreases at a fast rate even in the case of non-strongly-convex objective functions. Under the global regularity assumption, we provide a simple yet efficient algorithm called distributed randomized smoothing (DRS) based on a local smoothing of the objective function, and show that DRS is within a $d^{1/4}$ multiplicative factor of the optimal convergence rate, where d is the underlying dimension.

Graph Oracle Models, Lower Bounds, and Gaps for Parallel Stochastic Optimization

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #11

Blake Woodworth · Jialei Wang · Adam Smith · Brendan McMahan · Nati Srebro

We suggest a general oracle-based framework that captures parallel stochastic optimization in different parallelization settings described by a dependency graph, and derive generic lower bounds in terms of this graph. We then use the framework and derive lower bounds to study several specific parallel optimization settings, including delayed updates and parallel processing with intermittent communication. We highlight gaps between lower and upper bounds on the oracle complexity, and cases where the ``natural'' algorithms are not known to be optimal.

(Probably) Concave Graph Matching

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #12

Haggai Maron · Yaron Lipman

In this paper we address the graph matching problem. Following the recent works of \cite{zaslavskiy2009path,Vestner2017} we analyze and generalize the idea of concave relaxations. We introduce the concepts of \emph{conditionally concave} and \emph{probably conditionally concave} energies on polytopes and show that they encapsulate many instances of the graph matching problem, including matching Euclidean graphs and graphs on surfaces. We further prove that local minima of probably conditionally concave energies on general matching polytopes (\eg, doubly stochastic) are with high probability extreme points of the matching polytope (\eg, permutations).

Solving Non-smooth Constrained Programs with Lower Complexity than $\mathcal{O}(1/\varepsilon)$: A Primal-Dual Homotopy Smoothing Approach

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #13

Xiaohan Wei · Hao Yu · Qing Ling · Michael Neely

We propose a new primal-dual homotopy smoothing algorithm for a linearly constrained convex program, where neither the primal nor the dual function has to be smooth or strongly convex. The best known iteration complexity solving such a non-smooth problem is $\mathcal{O}(\varepsilon^{-1})$. In this paper, we show that by leveraging a local error bound condition on the dual function, the proposed algorithm can achieve a better primal convergence time of $\mathcal{O}\l(\varepsilon^{-2/(2+\beta)}\log_2(\varepsilon^{-1})\r)$, where $\beta\in(0,1]$ is a local error bound parameter. As an example application, we show that the distributed geometric median problem, which can be formulated as a constrained convex program, has its dual function non-smooth but satisfying the aforementioned local error bound condition with $\beta=1/2$, therefore enjoying a convergence time of $\mathcal{O}\l(\varepsilon^{-4/5}\log_2(\varepsilon^{-1})\r)$. This result improves upon the $\mathcal{O}(\varepsilon^{-1})$ convergence time bound achieved by existing distributed optimization algorithms. Simulation experiments also demonstrate the performance of our proposed algorithm.

Wasserstein Distributionally Robust Kalman Filtering

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #14

Soroosh Shafieezadeh Abadeh · Viet Anh Nguyen · Daniel Kuhn · Peyman Mohajerin Esfahani

We study a distributionally robust mean square error estimation problem over a nonconvex Wasserstein ambiguity set containing only normal distributions. We show that the optimal estimator and the least favorable distribution form a Nash equilibrium. Despite the non-convex nature of the ambiguity set, we prove that the estimation problem is equivalent to a tractable convex program. We further devise a Frank-Wolfe algorithm for this convex program whose direction-searching subproblem can be solved in a quasi-closed form. Using these ingredients, we introduce a distributionally robust Kalman filter that hedges against model risk.

Decentralize and Randomize: Faster Algorithm for Wasserstein Barycenters

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #15

Pavel Dvurechenskii · Darina Dvinskikh · Alexander Gasnikov · Cesar Uribe · Angelia Nedich

We study the decentralized distributed computation of discrete approximations for the regularized Wasserstein barycenter of a finite set of continuous probability measures distributedly stored over a network. We assume there is a network of agents/machines/computers, and each agent holds a private continuous probability measure and seeks to compute the barycenter of all the measures in the network by getting samples from its local measure and exchanging information with its neighbors. Motivated by this problem, we develop, and analyze, a novel accelerated primal-dual stochastic gradient method for general stochastic convex optimization problems with linear equality constraints. Then, we apply this method to the decen- tralized distributed optimization setting to obtain a new algorithm for the distributed semi-discrete regularized Wasserstein barycenter problem. Moreover, we show explicit non-asymptotic complexity for the proposed algorithm. Finally, we show the effectiveness of our method on the distributed computation of the regularized Wasserstein barycenter of univariate Gaussian and von Mises distributions, as well as some applications to image aggregation.

Limited Memory Kelley's Method Converges for Composite Convex and Submodular Objectives

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #16

Song Zhou · Swati Gupta · Madeleine Udell

The original simplicial method (OSM), a variant of the classic Kelley’s cutting plane method, has been shown to converge to the minimizer of a composite convex and submodular objective, though no rate of convergence for this method was known. Moreover, OSM is required to solve subproblems in each iteration whose size grows linearly in the number of iterations. We propose a limited memory version of Kelley’s method (L-KM) and of OSM that requires limited memory (at most n+ 1 constraints for an n-dimensional problem) independent of the iteration. We prove convergence for L-KM when the convex part of the objective g is strongly convex and show it converges linearly when g is also smooth. Our analysis relies on duality between minimization of the composite convex and submodular objective and minimization of a convex function over the submodular base polytope. We introduce a limited memory version, L-FCFW, of the Fully-Corrective Frank-Wolfe (FCFW) method with approximate correction, to solve the dual problem. We show that L-FCFW and L-KM are dual algorithms that produce the same sequence of iterates; hence both converge linearly (when g is smooth and strongly convex) and with limited memory. We propose L-KM to minimize composite convex and submodular objectives; however, our results on L-FCFW hold for general polytopes and may be of independent interest.

Stochastic Spectral and Conjugate Descent Methods

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #17

Dmitry Kovalev · Peter Richtarik · Eduard Gorbunov · Elnur Gasanov

The state-of-the-art methods for solving optimization problems in big dimensions are variants of randomized coordinate descent (RCD). In this paper we introduce a fundamentally new type of acceleration strategy for RCD based on the augmentation of the set of coordinate directions by a few spectral or conjugate directions. As we increase the number of extra directions to be sampled from, the rate of the method improves, and interpolates between the linear rate of RCD and a linear rate independent of the condition number. We develop and analyze also inexact variants of these methods where the spectral and conjugate directions are allowed to be approximate only. We motivate the above development by proving several negative results which highlight the limitations of RCD with importance sampling.

Third-order Smoothness Helps: Faster Stochastic Optimization Algorithms for Finding Local Minima

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #18

Yaodong Yu · Pan Xu · Quanquan Gu

We propose stochastic optimization algorithms that can find local minima faster than existing algorithms for nonconvex optimization problems, by exploiting the third-order smoothness to escape non-degenerate saddle points more efficiently. More specifically, the proposed algorithm only needs $\tilde{O}(\epsilon^{-10/3})$ stochastic gradient evaluations to converge to an approximate local minimum $\mathbf{x}$, which satisfies $\|\nabla f(\mathbf{x})\|_2\leq\epsilon$ and $\lambda_{\min}(\nabla^2 f(\mathbf{x}))\geq -\sqrt{\epsilon}$ in unconstrained stochastic optimization, where $\tilde{O}(\cdot)$ hides logarithm polynomial terms and constants. This improves upon the $\tilde{O}(\epsilon^{-7/2})$ gradient complexity achieved by the state-of-the-art stochastic local minima finding algorithms by a factor of $\tilde{O}(\epsilon^{-1/6})$. Experiments on two nonconvex optimization problems demonstrate the effectiveness of our algorithm and corroborate our theory.

First-order Stochastic Algorithms for Escaping From Saddle Points in Almost Linear Time

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #19

Yi Xu · Jing Rong · Tianbao Yang

(This is a theory paper) In this paper, we consider first-order methods for solving stochastic non-convex optimization problems. The key building block of the proposed algorithms is first-order procedures to extract negative curvature from the Hessian matrix through a principled sequence starting from noise, which are referred to {\it NEgative-curvature-Originated-from-Noise or NEON} and are of independent interest. Based on this building block, we design purely first-order stochastic algorithms for escaping from non-degenerate saddle points with a much better time complexity (almost linear time in the problem's dimensionality). In particular, we develop a general framework of {\it first-order stochastic algorithms} with a second-order convergence guarantee based on our new technique and existing algorithms that may only converge to a first-order stationary point. For finding a nearly {\it second-order stationary point} $\x$ such that $|\nabla F(\x)|\leq \epsilon$ and $\nabla^2 F(\x)\geq -\sqrt{\epsilon}I$ (in high probability), the best time complexity of the presented algorithms is $\widetilde O(d/\epsilon^{3.5})$, where $F(\cdot)$ denotes the objective function and d is the dimensionality of the problem. To the best of our knowledge, this is the first theoretical result of first-order stochastic algorithms with an almost linear time in terms of problem's dimensionality for finding second-order stationary points, which is even competitive with existing stochastic algorithms hinging on the second-order information.

Gen-Oja: Simple & Efficient Algorithm for Streaming Generalized Eigenvector Computation

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #20

Kush Bhatia · Aldo Pacchiano · Nicolas Flammarion · Peter Bartlett · Michael Jordan

In this paper, we study the problems of principle Generalized Eigenvector computation and Canonical Correlation Analysis in the stochastic setting. We propose a simple and efficient algorithm for these problems. We prove the global convergence of our algorithm, borrowing ideas from the theory of fast-mixing Markov chains and two-Time-Scale Stochastic Approximation, showing that it achieves the optimal rate of convergence. In the process, we develop tools for understanding stochastic processes with Markovian noise which might be of independent interest.

Sparse DNNs with Improved Adversarial Robustness

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #21

Yiwen Guo · Chao Zhang · Changshui Zhang · Yurong Chen

Deep neural networks (DNNs) are computationally/memory-intensive and vulnerable to adversarial attacks, making them prohibitive in some real-world applications. By converting dense models into sparse ones, pruning appears to be a promising solution to reducing the computation/memory cost. This paper studies classification models, especially DNN-based ones, to demonstrate that there exists intrinsic relationships between their sparsity and adversarial robustness. Our analyses reveal, both theoretically and empirically, that nonlinear DNN-based classifiers behave differently under l_2 attacks from some linear ones. We further demonstrate that an appropriately higher model sparsity implies better robustness of nonlinear DNNs, whereas over-sparsified models can be more difficult to resist adversarial examples.

Constructing Fast Network through Deconstruction of Convolution

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #22

Yunho Jeon · Junmo Kim

Convolutional neural networks have achieved great success in various vision tasks; however, they incur heavy resource costs. By using deeper and wider networks, network accuracy can be improved rapidly. However, in an environment with limited resources (e.g., mobile applications), heavy networks may not be usable. This study shows that naive convolution can be deconstructed into a shift operation and pointwise convolution. To cope with various convolutions, we propose a new shift operation called active shift layer (ASL) that formulates the amount of shift as a learnable function with shift parameters. This new layer can be optimized end-to-end through backpropagation and it can provide optimal shift values. Finally, we apply this layer to a light and fast network that surpasses existing state-of-the-art networks.

Learning Loop Invariants for Program Verification

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #23

Xujie Si · Hanjun Dai · Mukund Raghothaman · Mayur Naik · Le Song

A fundamental problem in program verification concerns inferring loop invariants. The problem is undecidable and even practical instances are challenging. Inspired by how human experts construct loop invariants, we propose a reasoning framework Code2Inv that constructs the solution by multi-step decision making and querying an external program graph memory block. By training with reinforcement learning, Code2Inv captures rich program features and avoids the need for ground truth solutions as supervision. Compared to previous learning tasks in domains with graph-structured data, it addresses unique challenges, such as a binary objective function and an extremely sparse reward that is given by an automated theorem prover only after the complete loop invariant is proposed. We evaluate Code2Inv on a suite of 133 benchmark problems and compare it to three state-of-the-art systems. It solves 106 problems compared to 73 by a stochastic search-based system, 77 by a heuristic search-based system, and 100 by a decision tree learning-based system. Moreover, the strategy learned can be generalized to new programs: compared to solving new instances from scratch, the pre-trained agent is more sample efficient in finding solutions.

Learning Libraries of Subroutines for Neurally–Guided Bayesian Program Induction

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #24

Kevin Ellis · Lucas Morales · Mathias Sablé-Meyer · Armando Solar-Lezama · Josh Tenenbaum

Successful approaches to program induction require a hand-engineered domain-specific language (DSL), constraining the space of allowed programs and imparting prior knowledge of the domain.

We contribute a program induction algorithm that learns a DSL while jointly training a neural network to efficiently search for programs in the learned DSL. We use our model to synthesize functions on lists, edit text, and solve symbolic regression problems, showing how the model learns a domain-specific library of program components for expressing solutions to problems in the domain.

Learning to Infer Graphics Programs from Hand-Drawn Images

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #25

Kevin Ellis · Daniel Ritchie · Armando Solar-Lezama · Josh Tenenbaum

We introduce a model that learns to convert simple hand drawings into graphics programs written in a subset of \LaTeX.~The model combines techniques from deep learning and program synthesis. We learn a convolutional neural network that proposes plausible drawing primitives that explain an image. These drawing primitives are a specification (spec) of what the graphics program needs to draw. We learn a model that uses program synthesis techniques to recover a graphics program from that spec. These programs have constructs like variable bindings, iterative loops, or simple kinds of conditionals. With a graphics program in hand, we can correct errors made by the deep network and extrapolate drawings.

Towards Understanding Learning Representations: To What Extent Do Different Neural Networks Learn the Same Representation

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #26

Liwei Wang · Lunjia Hu · Jiayuan Gu · Zhiqiang Hu · Yue Wu · Kun He · John Hopcroft

It is widely believed that learning good representations is one of the main reasons for the success of deep neural networks. Although highly intuitive, there is a lack of theory and systematic approach quantitatively characterizing what representations do deep neural networks learn. In this work, we move a tiny step towards a theory and better understanding of the representations. Specifically, we study a simpler problem: How similar are the representations learned by two networks with identical architecture but trained from different initializations. We develop a rigorous theory based on the neuron activation subspace match model. The theory gives a complete characterization of the structure of neuron activation subspace matches, where the core concepts are maximum match and simple match which describe the overall and the finest similarity between sets of neurons in two networks respectively. We also propose efficient algorithms to find the maximum match and simple matches. Finally, we conduct extensive experiments using our algorithms. Experimental results suggest that, surprisingly, representations learned by the same convolutional layers of networks trained from different initializations are not as similar as prevalently expected, at least in terms of subspace match.

Norm matters: efficient and accurate normalization schemes in deep networks

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #27

Elad Hoffer · Ron Banner · Itay Golan · Daniel Soudry

Over the past few years, Batch-Normalization has been commonly used in deep networks, allowing faster training and high performance for a wide variety of applications. However, the reasons behind its merits remained unanswered, with several shortcomings that hindered its use for certain tasks. In this work, we present a novel view on the purpose and function of normalization methods and weight-decay, as tools to decouple weights' norm from the underlying optimized objective. This property highlights the connection between practices such as normalization, weight decay and learning-rate adjustments. We suggest several alternatives to the widely used L^2 batch-norm, using normalization in L^1 and L^∞ spaces that can substantially improve numerical stability in low-precision implementations as well as provide computational and memory benefits. We demonstrate that such methods enable the first batch-norm alternative to work for half-precision implementations. Finally, we suggest a modification to weight-normalization, which improves its performance on large-scale tasks.

ResNet with one-neuron hidden layers is a Universal Approximator

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #28

Hongzhou Lin · Stefanie Jegelka

We demonstrate that a very deep ResNet with stacked modules that have one neuron per hidden layer and ReLU activation functions can uniformly approximate any Lebesgue integrable function in d dimensions, i.e. \ell_1(R^d). Due to the identity mapping inherent to ResNets, our network has alternating layers of dimension one and d. This stands in sharp contrast to fully connected networks, which are not universal approximators if their width is the input dimension d [21,11]. Hence, our result implies an increase in representational power for narrow deep networks by the ResNet architecture.

Hyperbolic Neural Networks

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #29

Octavian Ganea · Gary Becigneul · Thomas Hofmann

Hyperbolic spaces have recently gained momentum in the context of machine learning due to their high capacity and tree-likeliness properties. However, the representational power of hyperbolic geometry is not yet on par with Euclidean geometry, firstly because of the absence of corresponding hyperbolic neural network layers. Here, we bridge this gap in a principled manner by combining the formalism of Möbius gyrovector spaces with the Riemannian geometry of the Poincaré model of hyperbolic spaces. As a result, we derive hyperbolic versions of important deep learning tools: multinomial logistic regression, feed-forward and recurrent neural networks. This allows to embed sequential data and perform classification in the hyperbolic space. Empirically, we show that, even if hyperbolic optimization tools are limited, hyperbolic sentence embeddings either outperform or are on par with their Euclidean variants on textual entailment and noisy-prefix recognition tasks.

A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #30

Kimin Lee · Kibok Lee · Honglak Lee · Jinwoo Shin

Detecting test samples drawn sufficiently far away from the training distribution statistically or adversarially is a fundamental requirement for deploying a good classifier in many real-world machine learning applications. However, deep neural networks with the softmax classifier are known to produce highly overconfident posterior distributions even for such abnormal samples. In this paper, we propose a simple yet effective method for detecting any abnormal samples, which is applicable to any pre-trained softmax neural classifier. We obtain the class conditional Gaussian distributions with respect to (low- and upper-level) features of the deep models under Gaussian discriminant analysis, which result in a confidence score based on the Mahalanobis distance. While most prior methods have been evaluated for detecting either out-of-distribution or adversarial samples, but not both, the proposed method achieves the state-of-the-art performances for both cases in our experiments. Moreover, we found that our proposed method is more robust in harsh cases, e.g., when the training dataset has noisy labels or small number of samples. Finally, we show that the proposed method enjoys broader usage by applying it to class-incremental learning: whenever out-of-distribution samples are detected, our classification rule can incorporate new classes well without further training deep models.

Improving Neural Program Synthesis with Inferred Execution Traces

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #31

Richard Shin · Illia Polosukhin · Dawn Song

The task of program synthesis, or automatically generating programs that are consistent with a provided specification, remains a challenging task in artificial intelligence. As in other fields of AI, deep learning-based end-to-end approaches have made great advances in program synthesis. However, more so than other fields such as computer vision, program synthesis provides greater opportunities to explicitly exploit structured information such as execution traces, which contain a superset of the information input/output pairs. While they are highly useful for program synthesis, as execution traces are more difficult to obtain than input/output pairs, we use the insight that we can split the process into two parts: infer the trace from the input/output example, then infer the program from the trace. This simple modification leads to state-of-the-art results in program synthesis in the Karel domain, improving accuracy to 81.3% from the 77.12% of prior work.

Scaling the Poisson GLM to massive neural datasets through polynomial approximations

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #32

David Zoltowski · Jonathan W Pillow

Recent advances in recording technologies have allowed neuroscientists to record simultaneous spiking activity from hundreds to thousands of neurons in multiple brain regions. Such large-scale recordings pose a major challenge to existing statistical methods for neural data analysis. Here we develop highly scalable approximate inference methods for Poisson generalized linear models (GLMs) that require only a single pass over the data. Our approach relies on a recently proposed method for obtaining approximate sufficient statistics for GLMs using polynomial approximations [Huggins et al., 2017], which we adapt to the Poisson GLM setting. We focus on inference using quadratic approximations to nonlinear terms in the Poisson GLM log-likelihood with Gaussian priors, for which we derive closed-form solutions to the approximate maximum likelihood and MAP estimates, posterior distribution, and marginal likelihood. We introduce an adaptive procedure to select the polynomial approximation interval and show that the resulting method allows for efficient and accurate inference and regularization of high-dimensional parameters. We use the quadratic estimator to fit a fully-coupled Poisson GLM to spike train data recorded from 831 neurons across five regions of the mouse brain for a duration of 41 minutes, binned at 1 ms resolution. Across all neurons, this model is fit to over 2 billion spike count bins and identifies fine-timescale statistical dependencies between neurons within and across cortical and subcortical areas.

Discovery of Latent 3D Keypoints via End-to-end Geometric Reasoning

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #33

Supasorn Suwajanakorn · Noah Snavely · Jonathan Tompson · Mohammad Norouzi

This paper presents KeypointNet, an end-to-end geometric reasoning framework to learn an optimal set of category-specific keypoints, along with their detectors to predict 3D keypoints in a single 2D input image. We demonstrate this framework on 3D pose estimation task by proposing a differentiable pose objective that seeks the optimal set of keypoints for recovering the relative pose between two views of an object. Our network automatically discovers a consistent set of keypoints across viewpoints of a single object as well as across all object instances of a given object class. Importantly, we find that our end-to-end approach using no ground-truth keypoint annotations outperforms a fully supervised baseline using the same neural network architecture for the pose estimation task. The discovered 3D keypoints across the car, chair, and plane categories of ShapeNet are visualized at https://keypoints.github.io/

Learning to Reconstruct Shapes from Unseen Classes

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #34

Xiuming Zhang · Zhoutong Zhang · Chengkai Zhang · Josh Tenenbaum · Bill Freeman · Jiajun Wu

From a single image, humans are able to perceive the full 3D shape of an object by exploiting learned shape priors from everyday life. Contemporary single-image 3D reconstruction algorithms aim to solve this task in a similar fashion, but often end up with priors that are highly biased by training classes. Here we present an algorithm, Generalizable Reconstruction (GenRe), designed to capture more generic, class-agnostic shape priors. We achieve this with an inference network and training procedure that combine 2.5D representations of visible surfaces (depth and silhouette), spherical shape representations of both visible and non-visible surfaces, and 3D voxel-based representations, in a principled manner that exploits the causal structure of how 3D shapes give rise to 2D images. Experiments demonstrate that GenRe performs well on single-view shape reconstruction, and generalizes to diverse novel objects from categories not seen during training.

Using Trusted Data to Train Deep Networks on Labels

Corrupted by Severe Noise

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #35

Dan Hendrycks · Mantas Mazeika · Duncan Wilson · Kevin Gimpel

The growing importance of massive datasets with the advent of deep learning makes robustness to label noise a critical property for classifiers to have. Sources of label noise include automatic labeling for large datasets, non-expert labeling, and label corruption by data poisoning adversaries. In the latter case, corruptions may be arbitrarily bad, even so bad that a classifier predicts the wrong labels with high confidence. To protect against such sources of noise, we leverage the fact that a small set of clean labels is often easy to procure. We demonstrate that robustness to label noise up to severe strengths can be achieved by using a set of trusted data with clean labels, and propose a loss correction that utilizes trusted examples in a data-efficient manner to mitigate the effects of label noise on deep neural network classifiers. Across vision and natural language processing tasks, we experiment with various label noises at several strengths, and show that our method significantly outperforms existing methods.

A Reduction for Efficient LDA Topic Reconstruction

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #37

Matteo Almanza · Flavio Chierichetti · Alessandro Panconesi · Andrea Vattani

We present a novel approach for LDA (Latent Dirichlet Allocation) topic reconstruction. The main technical idea is to show that the distribution over the documents generated by LDA can be transformed into a distribution for a much simpler generative model in which documents are generated from {\em the same set of topics} but have a much simpler structure: documents are single topic and topics are chosen uniformly at random. Furthermore, this reduction is approximation preserving, in the sense that approximate distributions-- the only ones we can hope to compute in practice-- are mapped into approximate distribution in the simplified world. This opens up the possibility of efficiently reconstructing LDA topics in a roundabout way. Compute an approximate document distribution from the given corpus, transform it into an approximate distribution for the single-topic world, and run a reconstruction algorithm in the uniform, single topic world-- a much simpler task than direct LDA reconstruction. Indeed, we show the viability of the approach by giving very simple algorithms for a generalization of two notable cases that have been studied in the literature, p-separability and Gibbs sampling for matrix-like topics.

A Unified View of Piecewise Linear Neural Network

Verification

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #38

Rudy Bunel · Ilker Turkaslan · Philip Torr · Pushmeet Kohli · Pawan K Mudigonda

The success of Deep Learning and its potential use in many safety-critical applications has motivated research on formal verification of Neural Network (NN) models. Despite the reputation of learned NN models to behave as black boxes and the theoretical hardness of proving their properties, researchers have been successful in verifying some classes of models by exploiting their piecewise linear structure and taking insights from formal methods such as Satisifiability Modulo Theory. These methods are however still far from scaling to realistic neural networks. To facilitate progress on this crucial area, we make two key contributions. First, we present a unified framework that encompasses previous methods. This analysis results in the identification of new methods that combine the strengths of multiple existing approaches, accomplishing a speedup of two orders of magnitude compared to the previous state of the art. Second, we propose a new data set of benchmarks which includes a collection of previously released testcases. We use the benchmark to provide the first experimental comparison of existing algorithms and identify the factors impacting the hardness of verification problems.

Optimization over Continuous and Multi-dimensional Decisions with Observational Data

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #39

Dimitris Bertsimas · Christopher McCord

We consider the optimization of an uncertain objective over continuous and multi-dimensional decision spaces in problems in which we are only provided with observational data. We propose a novel algorithmic framework that is tractable, asymptotically consistent, and superior to comparable methods on example problems. Our approach leverages predictive machine learning methods and incorporates information on the uncertainty of the predicted outcomes for the purpose of prescribing decisions. We demonstrate the efficacy of our method on examples involving both synthetic and real data sets.

The Convergence of Sparsified Gradient Methods

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #40

Dan Alistarh · Torsten Hoefler · Mikael Johansson · Nikola Konstantinov · Sarit Khirirat · Cedric Renggli

Distributed training of massive machine learning models, in particular deep neural networks, via Stochastic Gradient Descent (SGD) is becoming commonplace. Several families of communication-reduction methods, such as quantization, large-batch methods, and gradient sparsification, have been proposed. To date, gradient sparsification methods--where each node sorts gradients by magnitude, and only communicates a subset of the components, accumulating the rest locally--are known to yield some of the largest practical gains. Such methods can reduce the amount of communication per step by up to \emph{three orders of magnitude}, while preserving model accuracy. Yet, this family of methods currently has no theoretical justification. This is the question we address in this paper. We prove that, under analytic assumptions, sparsifying gradients by magnitude with local error correction provides convergence guarantees, for both convex and non-convex smooth objectives, for data-parallel SGD. The main insight is that sparsification methods implicitly maintain bounds on the maximum impact of stale updates, thanks to selection by magnitude. Our analysis and empirical validation also reveal that these methods do require analytical conditions to converge well, justifying existing heuristics.

Estimating Learnability in the Sublinear Data Regime

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #41

Weihao Kong · Gregory Valiant

We consider the problem of estimating how well a model class is capable of fitting a distribution of labeled data. We show that it is often possible to accurately estimate this ``learnability'' even when given an amount of data that is too small to reliably learn any accurate model. Our first result applies to the setting where the data is drawn from a d-dimensional distribution with isotropic covariance, and the label of each datapoint is an arbitrary noisy function of the datapoint. In this setting, we show that with $O(\sqrt{d})$ samples, one can accurately estimate the fraction of the variance of the label that can be explained via the best linear function of the data. We extend these techniques to a binary classification, and show that the prediction error of the best linear classifier can be accurately estimated given $O(\sqrt{d})$ labeled samples. For comparison, in both the linear regression and binary classification settings, even if there is no noise in the labels, a sample size linear in the dimension, d, is required to \emph{learn} any function correlated with the underlying model. We further extend our estimation approach to the setting where the data distribution has an (unknown) arbitrary covariance matrix, allowing these techniques to be applied to settings where the model class consists of a linear function applied to a nonlinear embedding of the data. We demonstrate the practical viability of our approaches on synthetic and real data. This ability to estimate the explanatory value of a set of features (or dataset), even in the regime in which there is too little data to realize that explanatory value, may be relevant to the scientific and industrial settings for which data collection is expensive and there are many potentially relevant feature sets that could be collected.

Learning convex polytopes with margin

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #42

Lee-Ad Gottlieb · Eran Kaufman · Aryeh Kontorovich · Gabriel Nivasch

We present improved algorithm for properly learning convex polytopes in the realizable PAC setting from data with a margin. Our learning algorithm constructs a consistent polytope as an intersection of about t log t halfspaces with margins in time polynomial in t (where t is the number of halfspaces forming an optimal polytope). We also identify distinct generalizations of the notion of margin from hyperplanes to polytopes and investigate how they relate geometrically; this result may be of interest beyond the learning setting.

Ridge Regression and Provable Deterministic Ridge Leverage Score Sampling

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #43

Shannon McCurdy

Ridge leverage scores provide a balance between low-rank approximation and regularization, and are ubiquitous in randomized linear algebra and machine learning. Deterministic algorithms are also of interest in the moderately big data regime, because deterministic algorithms provide interpretability to the practitioner by having no failure probability and always returning the same results. We provide provable guarantees for deterministic column sampling using ridge leverage scores. The matrix sketch returned by our algorithm is a column subset of the original matrix, yielding additional interpretability. Like the randomized counterparts, the deterministic algorithm provides $(1+\epsilon)$ error column subset selection, $(1+\epsilon)$ error projection-cost preservation, and an additive-multiplicative spectral bound. We also show that under the assumption of power-law decay of ridge leverage scores, this deterministic algorithm is provably as accurate as randomized algorithms. Lastly, ridge regression is frequently used to regularize ill-posed linear least-squares problems. While ridge regression provides shrinkage for the regression coefficients, many of the coefficients remain small but non-zero. Performing ridge regression with the matrix sketch returned by our algorithm and a particular regularization parameter forces coefficients to zero and has a provable $(1+\epsilon)$ bound on the statistical risk. As such, it is an interesting alternative to elastic net regularization.

GIANT: Globally Improved Approximate Newton Method for Distributed Optimization

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #44

Shusen Wang · Farbod Roosta-Khorasani · Peng Xu · Michael W Mahoney

For distributed computing environment, we consider the empirical risk minimization problem and propose a distributed and communication-efficient Newton-type optimization method. At every iteration, each worker locally finds an Approximate NewTon (ANT) direction, which is sent to the main driver. The main driver, then, averages all the ANT directions received from workers to form a Globally Improved ANT (GIANT) direction. GIANT is highly communication efficient and naturally exploits the trade-offs between local computations and global communications in that more local computations result in fewer overall rounds of communications. Theoretically, we show that GIANT enjoys an improved convergence rate as compared with first-order methods and existing distributed Newton-type methods. Further, and in sharp contrast with many existing distributed Newton-type methods, as well as popular first-order methods, a highly advantageous practical feature of GIANT is that it only involves one tuning parameter. We conduct large-scale experiments on a computer cluster and, empirically, demonstrate the superior performance of GIANT.

Wavelet regression and additive models for irregularly spaced data

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #45

Asad Haris · Ali Shojaie · Noah Simon

We present a novel approach for nonparametric regression using wavelet basis functions. Our proposal, waveMesh, can be applied to non-equispaced data with sample size not necessarily a power of 2. We develop an efficient proximal gradient descent algorithm for computing the estimator and establish adaptive minimax convergence rates. The main appeal of our approach is that it naturally extends to additive and sparse additive models for a potentially large number of covariates. We prove minimax optimal convergence rates under a weak compatibility condition for sparse additive models. The compatibility condition holds when we have a small number of covariates. Additionally, we establish convergence rates for when the condition is not met. We complement our theoretical results with empirical studies comparing waveMesh to existing methods.

New Insight into Hybrid Stochastic Gradient Descent:

Beyond With-Replacement Sampling and Convexity

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #46

Pan Zhou · Xiaotong Yuan · Jiashi Feng

As an incremental-gradient algorithm, the hybrid stochastic gradient descent (HSGD) enjoys merits of both stochastic and full gradient methods for finite-sum minimization problem. However, the existing rate-of-convergence analysis for HSGD is made under with-replacement sampling (WRS) and is restricted to convex problems. It is not clear whether HSGD still carries these advantages under the common practice of without-replacement sampling (WoRS) for non-convex problems. In this paper, we affirmatively answer this open question by showing that under WoRS and for both convex and non-convex problems, it is still possible for HSGD (with constant step-size) to match full gradient descent in rate of convergence, while maintaining comparable sample-size-independent incremental first-order oracle complexity to stochastic gradient descent. For a special class of finite-sum problems with linear prediction models, our convergence results can be further improved in some cases. Extensive numerical results confirm our theoretical affirmation and demonstrate the favorable efficiency of WoRS-based HSGD.

Sample Efficient Stochastic Gradient Iterative Hard Thresholding Method for Stochastic Sparse Linear Regression with Limited Attribute Observation

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #47

Tomoya Murata · Taiji Suzuki

We develop new stochastic gradient methods for efficiently solving sparse linear regression in a partial attribute observation setting, where learners are only allowed to observe a fixed number of actively chosen attributes per example at training and prediction times. It is shown that the methods achieve essentially a sample complexity of $O(1/\varepsilon)$ to attain an error of ε under a variant of restricted eigenvalue condition, and the rate has better dependency on the problem dimension than existing methods. Particularly, if the smallest magnitude of the non-zero components of the optimal solution is not too small, the rate of our proposed {\it Hybrid} algorithm can be boosted to near the minimax optimal sample complexity of {\it full information} algorithms. The core ideas are (i) efficient construction of an unbiased gradient estimator by the iterative usage of the hard thresholding operator for configuring an exploration algorithm; and (ii) an adaptive combination of the exploration and an exploitation algorithms for quickly identifying the support of the optimum and efficiently searching the optimal parameter in its support. Experimental results are presented to validate our theoretical findings and the superiority of our proposed methods.

Robust Subspace Approximation in a Stream

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #48

Roie Levin · Anish Prasad Sevekari · David Woodruff

We study robust subspace estimation in the streaming and distributed settings. Given a set of n data points {a_i}*{i=1}^n in R^d and an integer k, we wish to find a linear subspace S of dimension k for which sum_i M(dist(S, a_i)) is minimized, where dist(S,x) := min*{y in S} |x-y|_2, and M() is some loss function. When M is the identity function, S gives a subspace that is more robust to outliers than that provided by the truncated SVD. Though the problem is NP-hard, it is approximable within a (1+epsilon) factor in polynomial time when k and epsilon are constant. We give the first sublinear approximation algorithm for this problem in the turnstile streaming and arbitrary partition distributed models, achieving the same time guarantees as in the offline case. Our algorithm is the first based entirely on oblivious dimensionality reduction, and significantly simplifies prior methods for this problem, which held in neither the streaming nor distributed models.

A Practical Algorithm for Distributed Clustering and Outlier Detection

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #50

Jiecao Chen · Erfan Sadeqi Azer · Qin Zhang

We study the classic k-means/median clustering, which are fundamental problems in unsupervised learning, in the setting where data are partitioned across multiple sites, and where we are allowed to discard a small portion of the data by labeling them as outliers. We propose a simple approach based on constructing small summary for the original dataset. The proposed method is time and communication efficient, has good approximation guarantees, and can identify the global outliers effectively. To the best of our knowledge, this is the first practical algorithm with theoretical guarantees for distributed clustering with outliers. Our experiments on both real and synthetic data have demonstrated the clear superiority of our algorithm against all the baseline algorithms in almost all metrics.

Compact Representation of Uncertainty in Clustering

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #51

Craig Greenberg · Nicholas Monath · Ari Kobren · Patrick Flaherty · Andrew McGregor · Andrew

McCallum

For many classic structured prediction problems, probability distributions over the dependent variables can be efficiently computed using widely-known algorithms and data structures (such as forward-backward, and its corresponding trellis for exact probability distributions in Markov models). However, we know of no previous work studying efficient representations of exact distributions over clusterings. This paper presents definitions and proofs for a dynamic-programming inference procedure that computes the partition function, the marginal probability of a cluster, and the MAP clustering---all exactly. Rather than the Nth Bell number, these exact solutions take time and space proportional to the substantially smaller powerset of N. Indeed, we improve upon the time complexity of the algorithm introduced by Kohonen and Corander (2016) for this problem by a factor of N. While still large, this previously unknown result is intellectually interesting in its own right, makes feasible exact inference for important real-world small data applications (such as medicine), and provides a natural stepping stone towards sparse-trellis approximations that enable further scalability (which we also explore). In experiments, we demonstrate the superiority of our approach over approximate methods in analyzing real-world gene expression data used in cancer treatment.

Bipartite Stochastic Block Models with Tiny Clusters

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #52

Stefan Neumann

We study the problem of finding clusters in random bipartite graphs. We present a simple two-step algorithm which provably finds even tiny clusters of size $O(n^\epsilon)$, where n is the number of vertices in the graph and $\epsilon > 0$. Previous algorithms were only able to identify clusters of size $\Omega(\sqrt{n})$. We evaluate the algorithm on synthetic and on real-world data; the experiments show that the algorithm can find extremely small clusters even in presence of high destructive noise.

Clustering Redemption–Beyond the Impossibility of Kleinberg’s Axioms

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #53

Vincent Cohen-Addad · Varun Kanade · Frederik Mallmann-Trenn

Kleinberg (2002) stated three axioms that any clustering procedure should satisfy and showed there is no clustering procedure that simultaneously satisfies all three. One of these, called the consistency axiom, requires that when the data is modified in a helpful way, i.e. if points in the same cluster are made more similar and those in different ones made less similar, the algorithm should output the same clustering. To circumvent this impossibility result, research has focused on considering clustering procedures that have a clustering quality measure (or a cost) and showing that a modification of Kleinberg's axioms that takes cost into account lead to feasible clustering procedures. In this work, we take a different approach, based on the observation that the consistency axiom fails to be satisfied when the "correct" number of clusters changes. We modify this axiom by making use of cost functions to determine the correct number of clusters, and require that consistency holds only if the number of clusters remains unchanged. We show that single linkage satisfies the modified axioms, and if the input is well-clusterable, some popular procedures such as k-means also satisfy the axioms, taking a step towards explaining the success of these objective functions for guiding the design of algorithms.

Understanding Regularized Spectral Clustering via Graph Conductance

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #54

Yilin Zhang · Karl Rohe

This paper uses the relationship between graph conductance and spectral clustering to study (i) the failures of spectral clustering and (ii) the benefits of regularization. The explanation is simple. Sparse and stochastic graphs create several dangling sets'', or small trees that are connected to the core of the graph by only one edge. Graph conductance is sensitive to these noisy dangling sets and spectral clustering inherits this sensitivity. The second part of the paper starts from a previously proposed form of regularized spectral clustering and shows that it is related to the graph conductance on aregularized graph''. When graph conductance is computed on the regularized graph, we call it CoreCut. Based upon previous arguments that relate graph conductance to spectral clustering (e.g. Cheeger inequality), minimizing CoreCut relaxes to regularized spectral clustering. Simple inspection of CoreCut reveals why it is less sensitive to dangling sets. Together, these results show that unbalanced partitions from spectral clustering can be understood as overfitting to noise in the periphery of a sparse and stochastic graph. Regularization fixes this overfitting. In addition to this statistical benefit, these results also demonstrate how regularization can improve the computational speed of spectral clustering. We provide simulations and data examples to illustrate these results.

Query K-means Clustering and the Double Dixie Cup Problem

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #55

I Chien · Chao Pan · Olgica Milenkovic

We consider the problem of approximate K-means clustering with outliers and side information provided by same-cluster queries and possibly noisy answers. Our solution shows that, under some mild assumptions on the smallest cluster size, one can obtain an $(1+\epsilon)$-approximation for the optimal potential with probability at least $1-\delta$, where $\epsilon>0$ and $\delta\in(0,1)$, using an expected number of $O(\frac{K^3}{\epsilon \delta})$ noiseless same-cluster queries and comparison-based clustering of complexity $O(ndK + \frac{K^3}{\epsilon \delta})$; here, n denotes the number of points and d the dimension of space. Compared to a handful of other known approaches that perform importance sampling to account for small cluster sizes, the proposed query technique reduces the number of queries by a factor of roughly $O(\frac{K^6}{\epsilon^3})$, at the cost of possibly missing very small clusters. We extend this settings to the case where some queries to the oracle produce erroneous information, and where certain points, termed outliers, do not belong to any clusters. Our proof techniques differ from previous methods used for K-means clustering analysis, as they rely on estimating the sizes of the clusters and the number of points needed for accurate centroid estimation and subsequent nontrivial generalizations of the double Dixie cup problem. We illustrate the performance of the proposed algorithm both on synthetic and real datasets, including MNIST and CIFAR 10.

How to tell when a clustering is (approximately) correct using convex relaxations

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #56

Marina Meila

We introduce the Sublevel Set (SS) method, a generic method to obtain sufficient guarantees of near-optimality and uniqueness (up to small perturbations) for a clustering. This method can be instantiated for a variety of clustering loss functions for which convex relaxations exist. Obtaining the guarantees in practice amounts to solving a convex optimization. We demonstrate the applicability of this method by obtaining distribution free guarantees for K-means clustering on realistic data sets.

Derivative Estimation in Random Design

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #57

Yu Liu · Kris De Brabanter

We propose a nonparametric derivative estimation method for random design without having to estimate the regression function. The method is based on a variance-reducing linear combination of symmetric difference quotients. First, we discuss the special case of uniform random design and establish the estimator’s asymptotic properties. Secondly, we generalize these results for any distribution of the dependent variable and compare the proposed estimator with popular estimators for derivative estimation such as local polynomial regression and smoothing splines.

Exploiting Numerical Sparsity for Efficient Learning : Faster Eigenvector Computation and Regression

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #58

Neha Gupta · Aaron Sidford

In this paper, we obtain improved running times for regression and top eigenvector computation for numerically sparse matrices. Given a data matrix $\mat{A} \in \R^{n \times d}$ where every row $a \in \R^d$ has $|a|_2^2 \leq L$ and numerical sparsity $\leq s$, i.e. $|a|_1^2 / |a|_2^2 \leq s$, we provide faster algorithms for these problems for many parameter settings.

For top eigenvector computation, when $\gap > 0$ is the relative gap between the top two eigenvectors of $\mat{A}^\top \mat{A}$ and r is the stable rank of $\mat{A}$ we obtain a running time of $\otilde(nd + r(s + \sqrt{r s}) / \gap^2)$ improving upon the previous best unaccelerated running time of $O(nd + r d / \gap^2)$. As $r \leq d$ and $s \leq d$ our algorithm everywhere improves or matches the previous bounds for all parameter settings.

For regression, when $\mu > 0$ is the smallest eigenvalue of $\mat{A}^\top \mat{A}$ we obtain a running time of $\otilde(nd + (nL / \mu) \sqrt{s nL / \mu})$ improving upon the previous best unaccelerated running time of $\otilde(nd + n L d / \mu)$. This result expands when regression can be solved in nearly linear time from when $L/\mu = \otilde(1)$ to when $L / \mu = \otilde(d^{2/3} / (sn)^{1/3})$.

Furthermore, we obtain similar improvements even when row norms and numerical sparsities are non-uniform and we show how to achieve even faster running times by accelerating using approximate proximal point \cite{frostig2015regularizing} / catalyst \cite{lin2015universal}. Our running times depend only on the size of the input and natural numerical measures of the matrix, i.e. eigenvalues and ℓ_p norms, making progress on a key open problem regarding optimal running times for efficient large-scale learning.

Boosted Sparse and Low-Rank Tensor Regression

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #59

Lifang He · Kun Chen · Wanwan Xu · Jiayu Zhou · Fei Wang

We propose a sparse and low-rank tensor regression model to relate a univariate outcome to a feature tensor, in which each unit-rank tensor from the CP decomposition of the coefficient tensor is assumed to be sparse. This structure is both parsimonious and highly interpretable, as it implies that the outcome is related to the features through a few distinct pathways, each of which may only involve subsets of feature dimensions. We take a divide-and-conquer strategy to simplify the task into a set of sparse unit-rank tensor regression problems. To make the computation efficient and scalable, for the unit-rank tensor regression, we propose a stagewise estimation procedure to efficiently trace out its entire solution path. We show that as the step size goes to zero, the stagewise solution paths converge exactly to those of the corresponding regularized regression. The superior performance of our approach is demonstrated on various real-world and synthetic examples.

An Efficient Pruning Algorithm for Robust Isotonic Regression

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #60

Cong Han Lim

We study a generalization of the classic isotonic regression problem where we allow separable nonconvex objective functions, focusing on the case of estimators used in robust regression. A simple dynamic programming approach allows us to solve this problem to within ε-accuracy (of the global minimum) in time linear in 1/ε and the dimension. We can combine techniques from the convex case with branch-and-bound ideas to form a new algorithm for this problem that naturally exploits the shape of the objective function. Our algorithm achieves the best bounds for both the general nonconvex and convex case (linear in log (1/ε)), while performing much faster in practice than a straightforward dynamic programming approach, especially as the desired accuracy increases.

A convex program for bilinear inversion of sparse vectors

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #61

Alireza Aghasi · Ali Ahmed · Paul Hand · Babhru Joshi

We consider the bilinear inverse problem of recovering two vectors, x in R^L and w in R^L, from their entrywise product. We consider the case where x and w have known signs and are sparse with respect to known dictionaries of size K and N, respectively. Here, K and N may be larger than, smaller than, or equal to L. We introduce L1-BranchHull, which is a convex program posed in the natural parameter space and does not require an approximate solution or initialization in order to be stated or solved. We study the case where x and w are S1- and S2-sparse with respect to a random dictionary, with the sparse vectors satisfying an effective sparsity condition, and present a recovery guarantee that depends on the number of measurements as L > Omega(S1+S2)(log(K+N))^2. Numerical experiments verify that the scaling constant in the theorem is not too large. One application of this problem is the sweep distortion removal task in dielectric imaging, where one of the signals is a nonnegative reflectivity, and the other signal lives in a known subspace, for example that given by dominant wavelet coefficients. We also introduce a variants of L1-BranchHull for the purposes of tolerating noise and outliers, and for the purpose of recovering piecewise constant signals. We provide an ADMM implementation of these variants and show they can extract piecewise constant behavior from real images.

Efficient Convex Completion of Coupled Tensors using Coupled Nuclear Norms

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #62

Kishan Wimalawarne · Hiroshi Mamitsuka

Coupled norms have emerged as a convex method to solve coupled tensor completion. A limitation with coupled norms is that they only induce low-rankness using the multilinear rank of coupled tensors. In this paper, we introduce a new set of coupled norms known as coupled nuclear norms by constraining the CP rank of coupled tensors. We propose new coupled completion models using the coupled nuclear norms as regularizers, which can be optimized using computationally efficient optimization methods. We derive excess risk bounds for proposed coupled completion models and show that proposed norms lead to better performance. Through simulation and real-data experiments, we demonstrate that proposed norms achieve better performance for coupled completion compared to existing coupled norms.

Support Recovery for Orthogonal Matching Pursuit: Upper and Lower bounds

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #63

Raghav Somani · Chirag Gupta · Prateek Jain · Praneeth Netrapalli

This paper studies the problem of sparse regression where the goal is to learn a sparse vector that best optimizes a given objective function. Under the assumption that the objective function satisfies restricted strong convexity (RSC), we analyze orthogonal matching pursuit (OMP), a greedy algorithm that is used heavily in applications, and obtain support recovery result as well as a tight generalization error bound for OMP. Furthermore, we obtain lower bounds for OMP, showing that both our results on support recovery and generalization error are tight up to logarithmic factors. To the best of our knowledge, these support recovery and generalization bounds are the first such matching upper and lower bounds (up to logarithmic factors) for {\em any} sparse regression algorithm under the RSC assumption.

Learning without the Phase: Regularized PhaseMax Achieves Optimal Sample Complexity

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #64

Fariborz Salehi · Ehsan Abbasi · Babak Hassibi

The problem of estimating an unknown signal, $\mathbf x_0\in \mathbb R^n$, from a vector $\mathbf y\in \mathbb R^m$ consisting of m magnitude-only measurements of the form $y_i=|\mathbf a_i\mathbf x_0|$, where $\mathbf a_i$'s are the rows of a known measurement matrix $\mathbf A$ is a classical problem known as phase retrieval. This problem arises when measuring the phase is costly or altogether infeasible. In many applications in machine learning, signal processing, statistics, etc., the underlying signal has certain structure (sparse, low-rank, finite alphabet, etc.), opening of up the possibility of recovering $\mathbf x_0$ from a number of measurements smaller than the ambient dimension, i.e., $m

On Controllable Sparse Alternatives to Softmax

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #65

Anirban Laha · Saneem Ahmed Chemmengath · Priyanka Agrawal · Mitesh Khapra · Karthik Sankaranarayanan · Harish Ramaswamy

Converting an n-dimensional vector to a probability distribution over n objects is a commonly used component in many machine learning tasks like multiclass classification, multilabel classification, attention mechanisms etc. For this, several probability mapping functions have been proposed and employed in literature such as softmax, sum-normalization, spherical softmax, and sparsemax, but there is very little understanding in terms how they relate with each other. Further, none of the above formulations offer an explicit control over the degree of sparsity. To address this, we develop a unified framework that encompasses all these formulations as special cases. This framework ensures simple closed-form solutions and existence of sub-gradients suitable for learning via backpropagation. Within this framework, we propose two novel sparse formulations, sparsegen-lin and sparsehourglass, that seek to provide a control over the degree of desired sparsity. We further develop novel convex loss functions that help induce the behavior of aforementioned formulations in the multilabel classification setting, showing improved performance. We also demonstrate empirically that the proposed formulations, when used to compute attention weights, achieve better or comparable performance on standard seq2seq tasks like neural machine translation and abstractive summarization.

Sparse PCA from Sparse Linear Regression

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #66

Guy Bresler · Sung Min Park · Madalina Persu

Sparse Principal Component Analysis (SPCA) and Sparse Linear Regression (SLR) have a wide range of applications and have attracted a tremendous amount of attention in the last two decades as canonical examples of statistical problems in high dimension. A variety of algorithms have been proposed for both SPCA and SLR, but an explicit connection between the two had not been made. We show how to efficiently transform a black-box solver for SLR into an algorithm for SPCA: assuming the SLR solver satisfies prediction error guarantees achieved by existing efficient algorithms such as those based on the Lasso, the SPCA algorithm derived from it achieves near state of the art guarantees for testing and for support recovery for the single spiked covariance model as obtained by the current best polynomial-time algorithms. Our reduction not only highlights the inherent similarity between the two problems, but also, from a practical standpoint, allows one to obtain a collection of algorithms for SPCA directly from known algorithms for SLR. We provide experimental results on simulated data comparing our proposed framework to other algorithms for SPCA.

Efficient Anomaly Detection via Matrix Sketching

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #67

Vatsal Sharan · Parikshit Gopalan · Udi Wieder

We consider the problem of finding anomalies in high-dimensional data using popular PCA based anomaly scores. The naive algorithms for computing these scores explicitly compute the PCA of the covariance matrix which uses space quadratic in the dimensionality of the data. We give the first streaming algorithms that use space that is linear or sublinear in the dimension. We prove general results showing that \emph{any} sketch of a matrix that satisfies a certain operator norm guarantee can be used to approximate these scores. We instantiate these results with powerful matrix sketching techniques such as Frequent Directions and random projections to derive efficient and practical algorithms for these problems, which we validate over real-world data sets. Our main technical contribution is to prove matrix perturbation inequalities for operators arising in the computation of these measures.

Dimensionality Reduction for Stationary Time Series via Stochastic Nonconvex Optimization

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #69

Minshuo Chen · Lin Yang · Mengdi Wang · Tuo Zhao

Stochastic optimization naturally arises in machine learning. Efficient algorithms with provable guarantees, however, are still largely missing, when the objective function is nonconvex and the data points are dependent. This paper studies this fundamental challenge through a streaming PCA problem for stationary time series data. Specifically, our goal is to estimate the principle component of time series data with respect to the covariance matrix of the stationary distribution. Computationally, we propose a variant of Oja's algorithm combined with downsampling to control the bias of the stochastic gradient caused by the data dependency. Theoretically, we quantify the uncertainty of our proposed stochastic algorithm based on diffusion approximations. This allows us to prove the asymptotic rate of convergence and further implies near optimal asymptotic sample complexity. Numerical experiments are provided to support our analysis.

Contrastive Learning from Pairwise Measurements

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #70

Yi Chen · Zhuoran Yang · Yuchen Xie · Princeton Zhaoran Wang

Learning from pairwise measurements naturally arises from many applications, such as rank aggregation, ordinal embedding, and crowdsourcing. However, most existing models and algorithms are susceptible to potential model misspecification. In this paper, we study a semiparametric model where the pairwise measurements follow a natural exponential family distribution with an unknown base measure. Such a semiparametric model includes various popular parametric models, such as the Bradley-Terry-Luce model and the paired cardinal model, as special cases. To estimate this semiparametric model without specifying the base measure, we propose a data augmentation technique to create virtual examples, which enables us to define a contrastive estimator. In particular, we prove that such a contrastive estimator is invariant to model misspecification within the natural exponential family, and moreover, attains the optimal statistical rate of convergence up to a logarithmic factor. We provide numerical experiments to corroborate our theory.

Deep Functional Dictionaries: Learning Consistent Semantic Structures on 3D Models from Functions

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #71

Minhyuk Sung · Hao Su · Ronald Yu · Leonidas J Guibas

Various 3D semantic attributes such as segmentation masks, geometric features, keypoints, and materials can be encoded as per-point probe functions on 3D geometries. Given a collection of related 3D shapes, we consider how to jointly analyze such probe functions over different shapes, and how to discover common latent structures using a neural network — even in the absence of any correspondence information. Our network is trained on point cloud representations of shape geometry and associated semantic functions on that point cloud. These functions express a shared semantic understanding of the shapes but are not coordinated in any way. For example, in a segmentation task, the functions can be indicator functions of arbitrary sets of shape parts, with the particular combination involved not known to the network. Our network is able to produce a small dictionary of basis functions for each shape, a dictionary whose span includes the semantic functions provided for that shape. Even though our shapes have independent discretizations and no functional correspondences are provided, the network is able to generate latent bases, in a consistent order, that reflect the shared semantic structure among the shapes. We demonstrate the effectiveness of our technique in various segmentation and keypoint selection applications.

Large Scale computation of Means and Clusters for Persistence Diagrams using Optimal Transport

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #72

Theo Lacombe · Marco Cuturi · Steve OUDOT

Persistence diagrams (PDs) are now routinely used to summarize the underlying topology of complex data. Despite several appealing properties, incorporating PDs in learning pipelines can be challenging because their natural geometry is not Hilbertian. Indeed, this was recently exemplified in a string of papers which show that the simple task of averaging a few PDs can be computationally prohibitive. We propose in this article a tractable framework to carry out standard tasks on PDs at scale, notably evaluating distances, estimating barycenters and performing clustering. This framework builds upon a reformulation of PD metrics as optimal transport (OT) problems. Doing so, we can exploit recent computational advances: the OT problem on a planar grid, when regularized with entropy, is convex can be solved in linear time using the Sinkhorn algorithm and convolutions. This results in scalable computations that can stream on GPUs. We demonstrate the efficiency of our approach by carrying out clustering with diagrams metrics on several thousands of PDs, a scale never seen before in the literature.

Representation Learning of Compositional Data

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #73

Marta Avalos · Richard Nock · Cheng Soon Ong · Julien Rouar · Ke Sun

We consider the problem of learning a low dimensional representation for compositional data. Compositional data consists of a collection of nonnegative data that sum to a constant value. Since the parts of the collection are statistically dependent, many standard tools cannot be directly applied. Instead, compositional data must be first transformed before analysis. Focusing on principal component analysis (PCA), we propose an approach that allows low dimensional representation learning directly from the original data. Our approach combines the benefits of the log-ratio transformation from compositional data analysis and exponential family PCA. A key tool in its derivation is a generalization of the scaled Bregman theorem, that relates the perspective transform of a Bregman divergence to the Bregman divergence of a perspective transform and a remainder conformal divergence. Our proposed approach includes a convenient surrogate (upper bound) loss of the exponential family PCA which has an easy to optimize form. We also derive the corresponding form for nonlinear autoencoders. Experiments on simulated data and microbiome data show the promise of our method.

How To Make the Gradients Small Stochastically: Even Faster Convex and Nonconvex SGD

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #74

Zeyuan Allen-Zhu

Stochastic gradient descent (SGD) gives an optimal convergence rate when minimizing convex stochastic objectives $f(x)$. However, in terms of making the gradients small, the original SGD does not give an optimal rate, even when $f(x)$ is convex.

If $f(x)$ is convex, to find a point with gradient norm ε, we design an algorithm SGD3 with a near-optimal rate $\tilde{O}(\varepsilon^{-2})$, improving the best known rate $O(\varepsilon^{-8/3})$. If $f(x)$ is nonconvex, to find its ε-approximate local minimum, we design an algorithm SGD5 with rate $\tilde{O}(\varepsilon^{-3.5})$, where previously SGD variants only achieve $\tilde{O}(\varepsilon^{-4})$. This is no slower than the best known stochastic version of Newton's method in all parameter regimes.

Statistical Optimality of Stochastic Gradient Descent on Hard Learning Problems through Multiple Passes

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #75

Loucas Pillaud-Vivien · Alessandro Rudi · Francis Bach

We consider stochastic gradient descent (SGD) for least-squares regression with potentially several passes over the data. While several passes have been widely reported to perform practically better in terms of predictive performance on unseen data, the existing theoretical analysis of SGD suggests that a single pass is statistically optimal. While this is true for low-dimensional easy problems, we show that for hard problems, multiple passes lead to statistically optimal predictions while single pass does not; we also show that in these hard models, the optimal number of passes over the data increases with sample size. In order to define the notion of hardness and show that our predictive performances are optimal, we consider potentially infinite-dimensional models and notions typically associated to kernel methods, namely, the decay of eigenvalues of the covariance matrix of the features and the complexity of the optimal predictor as measured through the covariance matrix. We illustrate our results on synthetic experiments with non-linear kernel methods and on a classical benchmark with a linear model.

Optimal Subsampling with Influence Functions

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #76

Daniel Ting · Eric Brochu

Subsampling is a common and often effective method to deal with the computational challenges of large datasets. However, for most statistical models, there is no well-motivated approach for drawing a non-uniform subsample. We show that the concept of an asymptotically linear estimator and the associated influence function leads to asymptotically optimal sampling probabilities for a wide class of popular models. This is the only tight optimality result for subsampling we are aware of as other methods only provide probabilistic error bounds or optimal rates. Furthermore, for linear regression models, which have well-studied procedures for non-uniform subsampling, we empirically show our optimal influence function based method outperforms previous approaches even when using approximations to the optimal probabilities.

Metric on Nonlinear Dynamical Systems with Perron-Frobenius Operators

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #77

Isao Ishikawa · Keisuke Fujii · Masahiro Ikeda · Yuka Hashimoto · Yoshinobu Kawahara

The development of a metric for structural data is a long-term problem in pattern recognition and machine learning. In this paper, we develop a general metric for comparing nonlinear dynamical systems that is defined with Perron-Frobenius operators in reproducing kernel Hilbert spaces. Our metric includes the existing fundamental metrics for dynamical systems, which are basically defined with principal angles between some appropriately-chosen subspaces, as its special cases. We also describe the estimation of our metric from finite data. We empirically illustrate our metric with an example of rotation dynamics in a unit disk in a complex plane, and evaluate the performance with real-world time-series data.

Random Feature Stein Discrepancies

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #78

Jonathan Huggins · Lester Mackey

Computable Stein discrepancies have been deployed for a variety of applications, ranging from sampler selection in posterior inference to approximate Bayesian inference to goodness-of-fit testing. Existing convergence-determining Stein discrepancies admit strong theoretical guarantees but suffer from a computational cost that grows quadratically in the sample size. While linear-time Stein discrepancies have been proposed for goodness-of-fit testing, they exhibit avoidable degradations in testing power—even when power is explicitly optimized. To address these shortcomings, we introduce feature Stein discrepancies (ΦSDs), a new family of quality measures that can be cheaply approximated using importance sampling. We show how to construct ΦSDs that provably determine the convergence of a sample to its target and develop high-accuracy approximations—random ΦSDs (RΦSDs)—which are computable in near-linear time. In our experiments with sampler selection for approximate posterior inference and goodness-of-fit testing, RΦSDs perform as well or better than quadratic-time KSDs while being orders of magnitude faster to compute.

Informative Features for Model Comparison

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #79

Wittawat Jitkrittum · Heishiro Kanagawa · Patsorn Sangkloy · James Hays · Bernhard Schölkopf · Arthur Gretton

Given two candidate models, and a set of target observations, we address the problem of measuring the relative goodness of fit of the two models. We propose two new statistical tests which are nonparametric, computationally efficient (runtime complexity is linear in the sample size), and interpretable. As a unique advantage, our tests can produce a set of examples (informative features) indicating the regions in the data domain where one model fits significantly better than the other. In a real-world problem of comparing GAN models, the test power of our new test matches that of the state-of-the-art test of relative goodness of fit, while being one order of magnitude faster.

On Fast Leverage Score Sampling and Optimal Learning

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #80

Alessandro Rudi · Daniele Calandriello · Luigi Carratino · Lorenzo Rosasco

Leverage score sampling provides an appealing way to perform approximate com- putations for large matrices. Indeed, it allows to derive faithful approximations with a complexity adapted to the problem at hand. Yet, performing leverage scores sampling is a challenge in its own right requiring further approximations. In this paper, we study the problem of leverage score sampling for positive definite ma- trices defined by a kernel. Our contribution is twofold. First we provide a novel algorithm for leverage score sampling and second, we exploit the proposed method in statistical learning by deriving a novel solver for kernel ridge regression. Our main technical contribution is showing that the proposed algorithms are currently the most efficient and accurate for these problems.

Persistence Fisher Kernel: A Riemannian Manifold Kernel for Persistence Diagrams

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #81

Tam Le · Makoto Yamada

Algebraic topology methods have recently played an important role for statistical analysis with complicated geometric structured data such as shapes, linked twist maps, and material data. Among them, \textit{persistent homology} is a well-known tool to extract robust topological features, and outputs as \textit{persistence diagrams} (PDs). However, PDs are point multi-sets which can not be used in machine learning algorithms for vector data. To deal with it, an emerged approach is to use kernel methods, and an appropriate geometry for PDs is an important factor to measure the similarity of PDs. A popular geometry for PDs is the \textit{Wasserstein metric}. However, Wasserstein distance is not \textit{negative definite}. Thus, it is limited to build positive definite kernels upon the Wasserstein distance \textit{without approximation}. In this work, we rely upon the alternative \textit{Fisher information geometry} to propose a positive definite kernel for PDs \textit{without approximation}, namely the Persistence Fisher (PF) kernel. Then, we analyze eigensystem of the integral operator induced by the proposed kernel for kernel machines. Based on that, we derive generalization error bounds via covering numbers and Rademacher averages for kernel machines with the PF kernel. Additionally, we show some nice properties such as stability and infinite divisibility for the proposed kernel. Furthermore, we also propose a linear time complexity over the number of points in PDs for an approximation of our proposed kernel with a bounded error. Throughout experiments with many different tasks on various benchmark datasets, we illustrate that the PF kernel compares favorably with other baseline kernels for PDs.

Learning Bounds for Greedy Approximation with Explicit Feature Maps from Multiple Kernels

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #82

Shahin Shahrampour · Vahid Tarokh

Nonlinear kernels can be approximated using finite-dimensional feature maps for efficient risk minimization. Due to the inherent trade-off between the dimension of the (mapped) feature space and the approximation accuracy, the key problem is to identify promising (explicit) features leading to a satisfactory out-of-sample performance. In this work, we tackle this problem by efficiently choosing such features from multiple kernels in a greedy fashion. Our method sequentially selects these explicit features from a set of candidate features using a correlation metric. We establish an out-of-sample error bound capturing the trade-off between the error in terms of explicit features (approximation error) and the error due to spectral properties of the best model in the Hilbert space associated to the combined kernel (spectral error). The result verifies that when the (best) underlying data model is sparse enough, i.e., the spectral error is negligible, one can control the test error with a small number of explicit features, that can scale poly-logarithmically with data. Our empirical results show that given a fixed number of explicit features, the method can achieve a lower test error with a smaller time cost, compared to the state-of-the-art in data-dependent random features.

RetGK: Graph Kernels based on Return Probabilities of Random Walks

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #83

Zhen Zhang · Mianzhi Wang · Yijian Xiang · Yan Huang · Arye Nehorai

Graph-structured data arise in wide applications, such as computer vision, bioinformatics, and social networks. Quantifying similarities among graphs is a fundamental problem. In this paper, we develop a framework for computing graph kernels, based on return probabilities of random walks. The advantages of our proposed kernels are that they can effectively exploit various node attributes, while being scalable to large datasets. We conduct extensive graph classification experiments to evaluate our graph kernels. The experimental results show that our graph kernels significantly outperform other state-of-the-art approaches in both accuracy and computational efficiency.

Nonparametric Density Estimation under Adversarial Losses

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #84

Shashank Singh · Ananya Uppal · Boyue Li · Chun-Liang Li · Manzil Zaheer · Barnabas Poczos

We study minimax convergence rates of nonparametric density estimation under a large class of loss functions called ``adversarial losses'', which, besides classical L^p losses, includes maximum mean discrepancy (MMD), Wasserstein distance, and total variation distance. These losses are closely related to the losses encoded by discriminator networks in generative adversarial networks (GANs). In a general framework, we study how the choice of loss and the assumed smoothness of the underlying density together determine the minimax rate. We also discuss implications for training GANs based on deep ReLU networks, and more general connections to learning implicit generative models in a minimax statistical sense.

Deep Homogeneous Mixture Models: Representation, Separation, and Approximation

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #85

Priyank Jaini · Pascal Poupart · Yaoliang Yu

At their core, many unsupervised learning models provide a compact representation of homogeneous density mixtures, but their similarities and differences are not always clearly understood. In this work, we formally establish the relationships among latent tree graphical models (including special cases such as hidden Markov models and tensorial mixture models), hierarchical tensor formats and sum-product networks. Based on this connection, we then give a unified treatment of exponential separation in \emph{exact} representation size between deep mixture architectures and shallow ones. In contrast, for \emph{approximate} representation, we show that the conditional gradient algorithm can approximate any homogeneous mixture within ϵ accuracy by combining $O(1/\epsilon^2)$ ``shallow'' architectures, where the hidden constant may decrease (exponentially) with respect to the depth. Our experiments on both synthetic and real datasets confirm the benefits of depth in density estimation.

Gaussian Process Conditional Density Estimation

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #86

Vincent Dutordoir · Hugh Salimbeni · James Hensman · Marc Deisenroth

Conditional Density Estimation (CDE) models deal with estimating conditional distributions. The conditions imposed on the distribution are the inputs of the model. CDE is a challenging task as there is a fundamental trade-off between model complexity, representational capacity and overfitting. In this work, we propose to extend the model's input with latent variables and use Gaussian processes (GP) to map this augmented input onto samples from the conditional distribution. Our Bayesian approach allows for the modeling of small datasets, but we also provide the machinery for it to be applied to big data using stochastic variational inference. Our approach can be used to model densities even in sparse data regions, and allows for sharing learned structure between conditions. We illustrate the effectiveness and wide-reaching applicability of our model on a variety of real-world problems, such as spatio-temporal density estimation of taxi drop-offs, non-Gaussian noise modeling, and few-shot learning on omniglot images.

Low-rank Interaction with Sparse Additive Effects Model for Large Data Frames

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #87

Geneviève Robin · Hoi-To Wai · Julie Josse · Olga Klopp · Eric Moulines

Many applications of machine learning involve the analysis of large data frames -- matrices collecting heterogeneous measurements (binary, numerical, counts, etc.) across samples -- with missing values. Low-rank models, as studied by Udell et al. (2016), are popular in this framework for tasks such as visualization, clustering and missing value imputation. Yet, available methods with statistical guarantees and efficient optimization do not allow explicit modeling of main additive effects such as row and column, or covariate effects. In this paper, we introduce a low-rank interaction and sparse additive effects (LORIS) model which combines matrix regression on a dictionary and low-rank design, to estimate main effects and interactions simultaneously. We provide statistical guarantees in the form of upper bounds on the estimation error of both components. Then, we introduce a mixed coordinate gradient descent (MCGD) method which provably converges sub-linearly to an optimal solution and is computationally efficient for large scale data sets. We show on simulated and survey data that the method has a clear advantage over current practices.

Mixture Matrix Completion

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #88

Daniel Pimentel-Alarcon

Completing a data matrix X has become an ubiquitous problem in modern data science, with motivations in recommender systems, computer vision, and networks inference, to name a few. One typical assumption is that X is low-rank. A more general model assumes that each column of X corresponds to one of several low-rank matrices. This paper generalizes these models to what we call mixture matrix completion (MMC): the case where each entry of X corresponds to one of several low-rank matrices. MMC is a more accurate model for recommender systems, and brings more flexibility to other completion and clustering problems. We make four fundamental contributions about this new model. First, we show that MMC is theoretically possible (well-posed). Second, we give its precise information-theoretic identifiability conditions. Third, we derive the sample complexity of MMC. Finally, we give a practical algorithm for MMC with performance comparable to the state-of-the-art for simpler related problems, both on synthetic and real data.

Multivariate Time Series Imputation with Generative Adversarial Networks

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #89

Yonghong Luo · Xiangrui Cai · Ying ZHANG · Jun Xu · Yuan xiaojie

Multivariate time series usually contain a large number of missing values, which hinders the application of advanced analysis methods on multivariate time series data. Conventional approaches to addressing the challenge of missing values, including mean/zero imputation, case deletion, and matrix factorization-based imputation, are all incapable of modeling the temporal dependencies and the nature of complex distribution in multivariate time series. In this paper, we treat the problem of missing value imputation as data generation. Inspired by the success of Generative Adversarial Networks (GAN) in image generation, we propose to learn the overall distribution of a multivariate time series dataset with GAN, which is further used to generate the missing values for each sample. Different from the image data, the time series data are usually incomplete due to the nature of data recording process. A modified Gate Recurrent Unit is employed in GAN to model the temporal irregularity of the incomplete time series. Experiments on two multivariate time series datasets show that the proposed model outperformed the baselines in terms of accuracy of imputation. Experimental results also showed that a simple model on the imputed data can achieve state-of-the-art results on the prediction tasks, demonstrating the benefits of our model in downstream applications.

Fully Understanding The Hashing Trick

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #90

Lior Kamma · Casper B. Freksen · Kasper Green Larsen

Feature hashing, also known as {\em the hashing trick}, introduced by Weinberger et al. (2009), is one of the key techniques used in scaling-up machine learning algorithms. Loosely speaking, feature hashing uses a random sparse projection matrix $A : \mathbb{R}^n \to \mathbb{R}^m$ (where $m \ll n$) in order to reduce the dimension of the data from n to m while approximately preserving the Euclidean norm. Every column of A contains exactly one non-zero entry, equals to either -1 or 1.

Weinberger et al. showed tail bounds on $|Ax|_2^2$. Specifically they showed that for every ε, δ, if $|x|_{\infty} / |x|_2$ is sufficiently small, and m is sufficiently large, then \begin{equation}\Pr[\; | \;|Ax|_2^2 - |x|_2^2\; | < \varepsilon |x|_2^2 \;] \ge 1 - \delta \;.\end{equation} These bounds were later extended by Dasgupta et al. (2010) and most recently refined by Dahlgaard et al. (2017), however, the true nature of the performance of this key technique, and specifically the correct tradeoff between the pivotal parameters $|x|_{\infty} / |x|_2, m, \varepsilon, \delta$ remained an open question.

We settle this question by giving tight asymptotic bounds on the exact tradeoff between the central parameters, thus providing a complete understanding of the performance of feature hashing. We complement the asymptotic bound with empirical data, which shows that the constants "hiding" in the asymptotic notation are, in fact, very close to 1, thus further illustrating the tightness of the presented bounds in practice.

Learning semantic similarity in a continuous space

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #91

Michel Deudon

We address the problem of learning semantic representation of questions to measure similarity between pairs as a continuous distance metric. Our work naturally extends Word Mover’s Distance (WMD) [1] by representing text documents as normal distributions instead of bags of embedded words. Our learned metric measures the dissimilarity between two questions as the minimum amount of distance the intent (hidden representation) of one question needs to "travel" to match the intent of another question. We first learn to repeat, reformulate questions to infer intents as normal distributions with a deep generative model [2] (variational auto encoder). Semantic similarity between pairs is then learned discriminatively as an optimal transport distance metric (Wasserstein 2) with our novel variational siamese framework. Among known models that can read sentences individually, our proposed framework achieves competitive results on Quora duplicate questions dataset. Our work sheds light on how deep generative models can approximate distributions (semantic representations) to effectively measure semantic similarity with meaningful distance metrics from Information Theory.

Bilevel learning of the Group Lasso structure

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #92

Jordan Frecon · Saverio Salzo · Massimiliano Pontil

Regression with group-sparsity penalty plays a central role in high-dimensional prediction problems. Most of existing methods require the group structure to be known a priori. In practice, this may be a too strong assumption, potentially hampering the effectiveness of the regularization method. To circumvent this issue, we present a method to estimate the group structure by means of a continuous bilevel optimization problem where the data is split into training and validation sets. Our approach relies on an approximation scheme where the lower level problem is replaced by a smooth dual forward-backward algorithm with Bregman distances. We provide guarantees regarding the convergence of the approximate procedure to the exact problem and demonstrate the well behaviour of the proposed method on synthetic experiments. Finally, a preliminary application to genes expression data is tackled with the purpose of unveiling functional groups.

Bayesian Structure Learning by Recursive Bootstrap

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #93

Raanan Y. Rohekar · Yaniv Gurwicz · Shami Nisimov · Guy Koren · Gal Novik

We address the problem of Bayesian structure learning for domains with hundreds of variables by employing non-parametric bootstrap, recursively. We propose a method that covers both model averaging and model selection in the same framework. The proposed method deals with the main weakness of constraint-based learning---sensitivity to errors in the independence tests---by a novel way of combining bootstrap with constraint-based learning. Essentially, we provide an algorithm for learning a tree, in which each node represents a scored CPDAG for a subset of variables and the level of the node corresponds to the maximal order of conditional independencies that are encoded in the graph. As higher order independencies are tested in deeper recursive calls, they benefit from more bootstrap samples, and therefore are more resistant to the curse-of-dimensionality. Moreover, the re-use of stable low order independencies allows greater computational efficiency. We also provide an algorithm for sampling CPDAGs efficiently from their posterior given the learned tree. That is, not from the full posterior, but from a reduced space of CPDAGs encoded in the learned tree. We empirically demonstrate that the proposed algorithm scales well to hundreds of variables, and learns better MAP models and more reliable causal relationships between variables, than other state-of-the-art-methods.

Learning from Group Comparisons: Exploiting Higher Order Interactions

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #94

Yao Li · Minhao Cheng · Kevin Fujii · Fushing Hsieh · Cho-Jui Hsieh

We study the problem of learning from group comparisons, with applications in predicting outcomes of sports and online games. Most of the previous works in this area focus on learning individual effects---they assume each player has an underlying score, and the ''ability'' of the team is modeled by the sum of team members' scores. Therefore, all the current approaches cannot model deeper interaction between team members: some players perform much better if they play together, and some players perform poorly together. In this paper, we propose a new model that takes the player-interaction effects into consideration. However, under certain circumstances, the total number of individuals can be very large, and number of player interactions grows quadratically, which makes learning intractable. In this case, we propose a latent factor model, and show that the sample complexity of our model is bounded under mild assumptions. Finally, we show that our proposed models have much better prediction power on several E-sports datasets, and furthermore can be used to reveal interesting patterns that cannot be discovered by previous methods.

A Structured Prediction Approach for Label Ranking

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #95

Anna Korba · Alexandre Garcia · Florence d'Alché-Buc

We propose to solve a label ranking problem as a structured output regression task. In this view, we adopt a least square surrogate loss approach that solves a supervised learning problem in two steps: a regression step in a well-chosen feature space and a pre-image (or decoding) step. We use specific feature maps/embeddings for ranking data, which convert any ranking/permutation into a vector representation. These embeddings are all well-tailored for our approach, either by resulting in consistent estimators, or by solving trivially the pre-image problem which is often the bottleneck in structured prediction. Their extension to the case of incomplete or partial rankings is also discussed. Finally, we provide empirical results on synthetic and real-world datasets showing the relevance of our method.

The Sample Complexity of Semi-Supervised Learning with Nonparametric Mixture Models

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #96

Chen Dan · Liu Leqi · Bryon Aragam · Pradeep Ravikumar · Eric Xing

We study the sample complexity of semi-supervised learning (SSL) and introduce new assumptions based on the mismatch between a mixture model learned from unlabeled data and the true mixture model induced by the (unknown) class conditional distributions. Under these assumptions, we establish an $\Omega(K\log K)$ labeled sample complexity bound without imposing parametric assumptions, where K is the number of classes. Our results suggest that even in nonparametric settings it is possible to learn a near-optimal classifier using only a few labeled samples. Unlike previous theoretical work which focuses on binary classification, we consider general multiclass classification ($K>2$), which requires solving a difficult permutation learning problem. This permutation defines a classifier whose classification error is controlled by the Wasserstein distance between mixing measures, and we provide finite-sample results characterizing the behaviour of the excess risk of this classifier. Finally, we describe three algorithms for computing these estimators based on a connection to bipartite graph matching, and perform experiments to illustrate the superiority of the MLE over the majority vote estimator.

Binary Classification from Positive-Confidence Data

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #97

Takashi Ishida · Gang Niu · Masashi Sugiyama

Can we learn a binary classifier from only positive data, without any negative data or unlabeled data? We show that if one can equip positive data with confidence (positive-confidence), one can successfully learn a binary classifier, which we name positive-confidence (Pconf) classification. Our work is related to one-class classification which is aimed at "describing" the positive class by clustering-related methods, but one-class classification does not have the ability to tune hyper-parameters and their aim is not on "discriminating" positive and negative classes. For the Pconf classification problem, we provide a simple empirical risk minimization framework that is model-independent and optimization-independent. We theoretically establish the consistency and an estimation error bound, and demonstrate the usefulness of the proposed method for training deep neural networks through experiments.

Learning SMaLL Predictors

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #98

Vikas Garg · Ofer Dekel · Lin Xiao

We introduce a new framework for learning in severely resource-constrained settings. Our technique delicately amalgamates the representational richness of multiple linear predictors with the sparsity of Boolean relaxations, and thereby yields classifiers that are compact, interpretable, and accurate. We provide a rigorous formalism of the learning problem, and establish fast convergence of the ensuing algorithm via relaxation to a minimax saddle point objective. We supplement the theoretical foundations of our work with an extensive empirical evaluation.

Contour location via entropy reduction leveraging multiple information sources

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #99

Alexandre Marques · Remi Lam · Karen Willcox

We introduce an algorithm to locate contours of functions that are expensive to evaluate. The problem of locating contours arises in many applications, including classification, constrained optimization, and performance analysis of mechanical and dynamical systems (reliability, probability of failure, stability, etc.). Our algorithm locates contours using information from multiple sources, which are available in the form of relatively inexpensive, biased, and possibly noisy approximations to the original function. Considering multiple information sources can lead to significant cost savings. We also introduce the concept of contour entropy, a formal measure of uncertainty about the location of the zero contour of a function approximated by a statistical surrogate model. Our algorithm locates contours efficiently by maximizing the reduction of contour entropy per unit cost.

Multi-Class Learning: From Theory to Algorithm

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #100

Jian Li · Yong Liu · Rong Yin · Hua Zhang · Lizhong Ding · Weiping Wang

In this paper, we study the generalization performance of multi-class classification and obtain a shaper data-dependent generalization error bound with fast convergence rate, substantially improving the state-of-art bounds in the existing data-dependent generalization analysis. The theoretical analysis motivates us to devise two effective multi-class kernel learning algorithms with statistical guarantees. Experimental results show that our proposed methods can significantly outperform the existing multi-class classification methods.

Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #101

Zhilu Zhang · Mert Sabuncu

Deep neural networks (DNNs) have achieved tremendous success in a variety of applications across many disciplines. Yet, their superior performance comes with the expensive cost of requiring correctly annotated large-scale datasets. Moreover, due to DNNs' rich capacity, errors in training labels can hamper performance. To combat this problem, mean absolute error (MAE) has recently been proposed as a noise-robust alternative to the commonly-used categorical cross entropy (CCE) loss. However, as we show in this paper, MAE can perform poorly with DNNs and large-scale datasets. Here, we present a theoretically grounded set of noise-robust loss functions that can be seen as a generalization of MAE and CCE. Proposed loss functions can be readily applied with any existing DNN architecture and algorithm, while yielding good performance in a wide range of noisy label scenarios. We report results from experiments conducted with CIFAR-10, CIFAR-100 and FASHION-MNIST datasets and synthetically generated noisy labels.

A Smoother Way to Train Structured Prediction Models

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #102

Venkata Krishna Pillutla · Vincent Roulet · Sham Kakade · Zaid Harchaoui

We present a framework to train a structured prediction model by performing smoothing on the inference algorithm it builds upon. Smoothing overcomes the non-smoothness inherent to the maximum margin structured prediction objective, and paves the way for the use of fast primal gradient-based optimization algorithms. We illustrate the proposed framework by developing a novel primal incremental optimization algorithm for the structural support vector machine. The proposed algorithm blends an extrapolation scheme for acceleration and an adaptive smoothing scheme and builds upon the stochastic variance-reduced gradient algorithm. We establish its worst-case global complexity bound and study several practical variants. We present experimental results on two real-world problems, namely named entity recognition and visual object localization. The experimental results show that the proposed framework allows us to build upon efficient inference algorithms to develop large-scale optimization algorithms for structured prediction which can achieve competitive performance on the two real-world problems.

Constrained Graph Variational Autoencoders for Molecule Design

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #103

Qi Liu · Miltiadis Allamanis · Marc Brockschmidt · Alexander Gaunt

Graphs are ubiquitous data structures for representing interactions between entities. With an emphasis on applications in chemistry, we explore the task of learning to generate graphs that conform to a distribution observed in training data. We propose a variational autoencoder model in which both encoder and decoder are graph-structured. Our decoder assumes a sequential ordering of graph extension steps and we discuss and analyze design choices that mitigate the potential downsides of this linearization. Experiments compare our approach with a wide range of baselines on the molecule generation task and show that our method is successful at matching the statistics of the original dataset on semantically important metrics. Furthermore, we show that by using appropriate shaping of the latent space, our model allows us to design molecules that are (locally) optimal in desired properties.

Learning Beam Search Policies via Imitation Learning

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #104

Renato Negrinho · Matthew Gormley · Geoffrey Gordon

Beam search is widely used for approximate decoding in structured prediction problems. Models often use a beam at test time but ignore its existence at train time, and therefore do not explicitly learn how to use the beam. We develop an unifying meta-algorithm for learning beam search policies using imitation learning. In our setting, the beam is part of the model and not just an artifact of approximate decoding. Our meta-algorithm captures existing learning algorithms and suggests new ones. It also lets us show novel no-regret guarantees for learning beam search policies.

Loss Functions for Multiset Prediction

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #105

Sean Welleck · Zixin Yao · Yu Gai · Jialin Mao · Zheng Zhang · Kyunghyun Cho

We study the problem of multiset prediction. The goal of multiset prediction is to train a predictor that maps an input to a multiset consisting of multiple items. Unlike existing problems in supervised learning, such as classification, ranking and sequence generation, there is no known order among items in a target multiset, and each item in the multiset may appear more than once, making this problem extremely challenging. In this paper, we propose a novel multiset loss function by viewing this problem from the perspective of sequential decision making. The proposed multiset loss function is empirically evaluated on two families of datasets, one synthetic and the other real, with varying levels of difficulty, against various baseline loss functions including reinforcement learning, sequence, and aggregated distribution matching loss functions. The experiments reveal the effectiveness of the proposed loss function over the others.

Learning Confidence Sets using Support Vector Machines

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #106

Wenbo Wang · Xingye Qiao

The goal of confidence-set learning in the binary classification setting is to construct two sets, each with a specific probability guarantee to cover a class. An observation outside the overlap of the two sets is deemed to be from one of the two classes, while the overlap is an ambiguity region which could belong to either class. Instead of plug-in approaches, we propose a support vector classifier to construct confidence sets in a flexible manner. Theoretically, we show that the proposed learner can control the non-coverage rates and minimize the ambiguity with high probability. Efficient algorithms are developed and numerical studies illustrate the effectiveness of the proposed method.

Fast Similarity Search via Optimal Sparse Lifting

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #107

Wenye Li · Jingwei Mao · Yin Zhang · Shuguang Cui

Similarity search is a fundamental problem in computing science with various applications and has attracted significant research attention, especially in large-scale search with high dimensions. Motivated by the evidence in biological science, our work develops a novel approach for similarity search. Fundamentally different from existing methods that typically reduce the dimension of the data to lessen the computational complexity and speed up the search, our approach projects the data into an even higher-dimensional space while ensuring the sparsity of the data in the output space, with the objective of further improving precision and speed. Specifically, our approach has two key steps. Firstly, it computes the optimal sparse lifting for given input samples and increases the dimension of the data while approximately preserving their pairwise similarity. Secondly, it seeks the optimal lifting operator that maps input samples to the optimal sparse lifting. Computationally, both steps are modeled as optimization problems that can be efficiently and effectively solved by the Frank-Wolfe algorithm. Simple as it is, our approach reported significantly improved results in empirical evaluations, and exhibited its high potentials in solving practical problems.

The Sparse Manifold Transform

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #108

Yubei Chen · Dylan Paiton · Bruno Olshausen

We present a signal representation framework called the sparse manifold transform that combines key ideas from sparse coding, manifold learning, and slow feature analysis. It turns non-linear transformations in the primary sensory signal space into linear interpolations in a representational embedding space while maintaining approximate invertibility. The sparse manifold transform is an unsupervised and generative framework that explicitly and simultaneously models the sparse discreteness and low-dimensional manifold structure found in natural scenes. When stacked, it also models hierarchical composition. We provide a theoretical description of the transform and

demonstrate properties of the learned representation on both synthetic data and natural videos.

When do random forests fail?

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #109

Cheng Tang · Damien Garreau · Ulrike von Luxburg

Random forests are learning algorithms that build large collections of random trees and make predictions by averaging the individual tree predictions. In this paper, we consider various tree constructions and examine how the choice of parameters affects the generalization error of the resulting random forests as the sample size goes to infinity. We show that subsampling of data points during the tree construction phase is important: Forests can become inconsistent with either no subsampling or too severe subsampling. As a consequence, even highly randomized trees can lead to inconsistent forests if no subsampling is used, which implies that some of the commonly used setups for random forests can be inconsistent. As a second consequence we can show that trees that have good performance in nearest-neighbor search can be a poor choice for random forests.

Diverse Ensemble Evolution: Curriculum Data-Model Marriage

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #110

Tianyi Zhou · Shengjie Wang · Jeff Bilmes

We study a new method (``Diverse Ensemble Evolution (DivE^2)'') to train an ensemble of machine learning models that assigns data to models at each training epoch based on each model's current expertise and an intra- and inter-model diversity reward. DivE^2 schedules, over the course of training epochs, the relative importance of these characteristics; it starts by selecting easy samples for each model, and then gradually adjusts towards the models having specialized and complementary expertise on subsets of the training data, thereby encouraging high accuracy of the ensemble. We utilize an intra-model diversity term on data assigned to each model, and an inter-model diversity term on data assigned to pairs of models, to penalize both within-model and cross-model redundancy. We formulate the data-model marriage problem as a generalized bipartite matching, represented as submodular maximization subject to two matroid constraints. DivE^2 solves a sequence of continuous-combinatorial optimizations with slowly varying objectives and constraints. The combinatorial part handles the data-model marriage while the continuous part updates model parameters based on the assignments. In experiments, DivE^2 outperforms other ensemble training methods under a variety of model aggregation techniques, while also maintaining competitive efficiency.

The Pessimistic Limits and Possibilities of Margin-based Losses in Semi-supervised Learning

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #111

Jesse Krijthe · Marco Loog

Consider a classification problem where we have both labeled and unlabeled data available. We show that for linear classifiers defined by convex margin-based surrogate losses that are decreasing, it is impossible to construct \emph{any} semi-supervised approach that is able to guarantee an improvement over the supervised classifier measured by this surrogate loss on the labeled and unlabeled data. For convex margin-based loss functions that also increase, we demonstrate safe improvements \emph{are} possible.

Semi-Supervised Learning with Declaratively Specified Entropy Constraints

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #112

Haitian Sun · William Cohen · Lidong Bing

We propose a technique for declaratively specifying strategies for semi-supervised learning (SSL). SSL methods based on different assumptions perform differently on different tasks, which leads to difficulties applying them in practice. In this paper, we propose to use entropy to unify many types of constraints. Our method can be used to easily specify ensembles of semi-supervised learners, as well as agreement constraints and entropic regularization constraints between these learners, and can be used to model both well-known heuristics such as co-training, and novel domain-specific heuristics. Besides, our model is flexible as to the underlying learning mechanism. Compared to prior frameworks for specifying SSL techniques, our technique achieves consistent improvements on a suite of well-studied SSL benchmarks, and obtains a new state-of-the-art result on a difficult relation extraction task.

Learning to Multitask

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #113

Yu Zhang · Ying Wei · Qiang Yang

Multitask learning has shown promising performance in many applications and many multitask models have been proposed. In order to identify an effective multitask model for a given multitask problem, we propose a learning framework called Learning to MultiTask (L2MT). To achieve the goal, L2MT exploits historical multitask experience which is organized as a training set consisting of several tuples, each of which contains a multitask problem with multiple tasks, a multitask model, and the relative test error. Based on such training set, L2MT first uses a proposed layerwise graph neural network to learn task embeddings for all the tasks in a multitask problem and then learns an estimation function to estimate the relative test error based on task embeddings and the representation of the multitask model based on a unified formulation. Given a new multitask problem, the estimation function is used to identify a suitable multitask model. Experiments on benchmark datasets show the effectiveness of the proposed L2MT framework.

CatBoost: unbiased boosting with categorical features

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #114

Liudmila Prokhorenkova · Gleb Gusev · Aleksandr Vorobev · Anna Veronika Dorogush · Andrey Gulin

This paper presents the key algorithmic techniques behind CatBoost, a new gradient boosting toolkit. Their combination leads to CatBoost outperforming other publicly available boosting implementations in terms of quality on a variety of datasets. Two critical algorithmic advances introduced in CatBoost are the implementation of ordered boosting, a permutation-driven alternative to the classic algorithm, and an innovative algorithm for processing categorical features. Both techniques were created to fight a prediction shift caused by a special kind of target leakage present in all currently existing implementations of gradient boosting algorithms. In this paper, we provide a detailed analysis of this problem and demonstrate that proposed algorithms solve it effectively, leading to excellent empirical results.

Supervised autoencoders: Improving generalization performance with unsupervised regularizers

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #115

Lei Le · Andrew Patterson · Martha White

Generalization performance is a central goal in machine learning, particularly when learning representations with large neural networks. A common strategy to improve generalization has been through the use of regularizers, typically as a norm constraining the parameters. Regularizing hidden layers in a neural network architecture, however, is not straightforward. There have been a few effective layer-wise suggestions, but without theoretical guarantees for improved performance. In this work, we theoretically and empirically analyze one such model, called a supervised auto-encoder: a neural network that predicts both inputs (reconstruction error) and targets jointly. We provide a novel generalization result for linear auto-encoders, proving uniform stability based on the inclusion of the reconstruction error---particularly as an improvement on simplistic regularization such as norms or even on more advanced regularizations such as the use of auxiliary tasks. Empirically, we then demonstrate that, across an array of architectures with a different number of hidden units and activation functions, the supervised auto-encoder compared to the corresponding standard neural network never harms performance and can significantly improve generalization.

Representation Learning for Treatment Effect Estimation from Observational Data

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #116

Liuyi Yao · Sheng Li · Yaliang Li · Mengdi Huai · Jing Gao · Aidong Zhang

Estimating individual treatment effect (ITE) is a challenging problem in causal inference, due to the missing counterfactuals and the selection bias. Existing ITE estimation methods mainly focus on balancing the distributions of control and treated groups, but ignore the local similarity information that is helpful. In this paper, we propose a local similarity preserved individual treatment effect (SITE) estimation method based on deep representation learning. SITE preserves local similarity and balances data distributions simultaneously, by focusing on several hard samples in each mini-batch. Experimental results on synthetic and three real-world datasets demonstrate the advantages of the proposed SITE method, compared with the state-of-the-art ITE estimation methods.

SimplE Embedding for Link Prediction in Knowledge Graphs

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #117

Seyed Mehran Kazemi · David Poole

Knowledge graphs contain knowledge about the world and provide a structured representation of this knowledge. Current knowledge graphs contain only a small subset of what is true in the world. Link prediction approaches aim at predicting new links for a knowledge graph given the existing links among the entities. Tensor factorization approaches have proved promising for such link prediction problems. Proposed in 1927, Canonical Polyadic (CP) decomposition is among the first tensor factorization approaches. CP generally performs poorly for link prediction as it learns two independent embedding vectors for each entity, whereas they are really tied. We present a simple enhancement of CP (which we call SimplE) to allow the two embeddings of each entity to be learned dependently. The complexity of SimplE grows linearly with the size of embeddings. The embeddings learned through SimplE are interpretable, and certain types of background knowledge can be incorporated into these embeddings through weight tying. We prove SimplE is fully expressive and derive a bound on the size of its embeddings for full expressivity. We show empirically that, despite its simplicity, SimplE outperforms several state-of-the-art tensor factorization techniques. SimplE's code is available on GitHub at https://github.com/Mehran-k/SimplE.

DeepProbLog: Neural Probabilistic Logic Programming

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #118

Robin Manhaeve · Sebastijan Dumancic · Angelika Kimmig · Thomas Demeester · Luc De Raedt

We introduce DeepProbLog, a probabilistic logic programming language that incorporates deep learning by means of neural predicates. We show how existing inference and learning techniques can be adapted for the new language. Our experiments demonstrate that DeepProbLog supports (i) both symbolic and subsymbolic representations and inference, (ii) program induction, (iii) probabilistic (logic) programming, and (iv) (deep) learning from examples. To the best of our knowledge, this work is the first to propose a framework where general-purpose neural networks and expressive probabilistic-logical modeling and reasoning are integrated in a way that exploits the full expressiveness and strengths of both worlds and can be trained end-to-end based on examples.

Watch Your Step: Learning Node Embeddings via Graph Attention

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #119

Sami Abu-El-Haija · Bryan Perozzi · Rami Al-Rfou · Alexander Alemi

Graph embedding methods represent nodes in a continuous vector space, preserving different types of relational information from the graph. There are many hyper-parameters to these methods (e.g. the length of a random walk) which have to be manually tuned for every graph. In this paper, we replace previously fixed hyper-parameters with trainable ones that we automatically learn via backpropagation. In particular, we propose a novel attention model on the power series of the transition matrix, which guides the random walk to optimize an upstream objective. Unlike previous approaches to attention models, the method that we propose utilizes attention parameters exclusively on the data itself (e.g. on the random walk), and are not used by the model for inference. We experiment on link prediction tasks, as we aim to produce embeddings that best-preserve the graph structure, generalizing to unseen information. We improve state-of-the-art results on a comprehensive suite of real-world graph datasets including social, collaboration, and biological networks, where we observe that our graph attention model can reduce the error by up to 20\%-40\%. We show that our automatically-learned attention parameters can vary significantly per graph, and correspond to the optimal choice of hyper-parameter if we manually tune existing methods.

Invariant Representations without Adversarial Training

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #120

Daniel Moyer · Shuyang Gao · Rob Brekelmans · Aram Galstyan · Greg Ver Steeg

Representations of data that are invariant to changes in specified factors are useful for a wide range of problems: removing potential biases in prediction problems, controlling the effects of covariates, and disentangling meaningful factors of variation. Unfortunately, learning representations that exhibit invariance to arbitrary nuisance factors yet remain useful for other tasks is challenging. Existing approaches cast the trade-off between task performance and invariance in an adversarial way, using an iterative minimax optimization. We show that adversarial training is unnecessary and sometimes counter-productive; we instead cast invariant representation learning as a single information-theoretic objective that can be directly optimized. We demonstrate that this approach matches or exceeds performance of state-of-the-art adversarial approaches for learning fair representations and for generative modeling with controllable transformations.

Domain-Invariant Projection Learning for Zero-Shot Recognition

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #121

An Zhao · Mingyu Ding · Jiechao Guan · Zhiwu Lu · Tao Xiang · Ji-Rong Wen

Zero-shot learning (ZSL) aims to recognize unseen object classes without any training samples, which can be regarded as a form of transfer learning from seen classes to unseen ones. This is made possible by learning a projection between a feature space and a semantic space (e.g. attribute space). Key to ZSL is thus to learn a projection function that is robust against the often large domain gap between the seen and unseen classes. In this paper, we propose a novel ZSL model termed domain-invariant projection learning (DIPL). Our model has two novel components: (1) A domain-invariant feature self-reconstruction task is introduced to the seen/unseen class data, resulting in a simple linear formulation that casts ZSL into a min-min optimization problem. Solving the problem is non-trivial, and a novel iterative algorithm is formulated as the solver, with rigorous theoretic algorithm analysis provided. (2) To further align the two domains via the learned projection, shared semantic structure among seen and unseen classes is explored via forming superclasses in the semantic space. Extensive experiments show that our model outperforms the state-of-the-art alternatives by significant margins.

Unsupervised Learning of View-invariant Action Representations

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #122

Junnan Li · Yongkang Wong · Qi Zhao · Mohan Kankanhalli

The recent success in human action recognition with deep learning methods mostly adopt the supervised learning paradigm, which requires significant amount of manually labeled data to achieve good performance. However, label collection is an expensive and time-consuming process. In this work, we propose an unsupervised learning framework, which exploits unlabeled data to learn video representations. Different from previous works in video representation learning, our unsupervised learning task is to predict 3D motion in multiple target views using video representation from a source view. By learning to extrapolate cross-view motions, the representation can capture view-invariant motion dynamics which is discriminative for the action. In addition, we propose a view-adversarial training method to enhance learning of view-invariant features. We demonstrate the effectiveness of the learned representations for action recognition on multiple datasets.

Neural Architecture Optimization

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #123

Renqian Luo · Fei Tian · Tao Qin · Enhong Chen · Tie-Yan Liu

Automatic neural architecture design has shown its potential in discovering powerful neural network architectures. Existing methods, no matter based on reinforcement learning or evolutionary algorithms (EA), conduct architecture search in a discrete space, which is highly inefficient. In this paper, we propose a simple and efficient method to automatic neural architecture design based on continuous optimization. We call this new approach neural architecture optimization (NAO). There are three key components in our proposed approach: (1) An encoder embeds/maps neural network architectures into a continuous space. (2) A predictor takes the continuous representation of a network as input and predicts its accuracy. (3) A decoder maps a continuous representation of a network back to its architecture. The performance predictor and the encoder enable us to perform gradient based optimization in the continuous space to find the embedding of a new architecture with potentially better accuracy. Such a better embedding is then decoded to a network by the decoder. Experiments show that the architecture discovered by our method is very competitive for image classification task on CIFAR-10 and language modeling task on PTB, outperforming or on par with the best results of previous architecture search methods with a significantly reduction of computational resources. Specifically we obtain 2.11% test set error rate for CIFAR-10 image classification task and 56.0 test set perplexity of PTB language modeling task. The best discovered architectures on both tasks are successfully transferred to other tasks such as CIFAR-100 and

WikiText-2. Furthermore, combined with the recent proposed weight sharing mechanism, we discover powerful architecture on CIFAR-10 (with error rate 3.53%) and on PTB (with test set perplexity 56.6), with very limited computational resources (less than 10 GPU hours) for both tasks.

Scalable Hyperparameter Transfer Learning

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #124

Valerio Perrone · Rodolphe Jenatton · Matthias W Seeger · Cedric Archambeau

Bayesian optimization (BO) is a model-based approach for gradient-free black-box function optimization, such as hyperparameter optimization. Typically, BO relies on conventional Gaussian process (GP) regression, whose algorithmic complexity is cubic in the number of evaluations. As a result, GP-based BO cannot leverage large numbers of past function evaluations, for example, to warm-start related BO runs. We propose a multi-task adaptive Bayesian linear regression model for transfer learning in BO, whose complexity is linear in the function evaluations: one Bayesian linear regression model is associated to each black-box function optimization problem (or task), while transfer learning is achieved by coupling the models through a shared deep neural net. Experiments show that the neural net learns a representation suitable for warm-starting the black-box optimization problems and that BO runs can be accelerated when the target black-box function (e.g., validation loss) is learned together with other related signals (e.g., training loss). The proposed method was found to be at least one order of magnitude faster that methods recently published in the literature.

Learning To Learn Around A Common Mean

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #125

Giulia Denevi · Carlo Ciliberto · Dimitris Stamos · Massimiliano Pontil

The problem of learning-to-learn (LTL) or meta-learning is gaining increasing attention due to recent empirical evidence of its effectiveness in applications. The goal addressed in LTL is to select an algorithm that works well on tasks sampled from a meta-distribution. In this work, we consider the family of algorithms given by a variant of Ridge Regression, in which the regularizer is the square distance to an unknown mean vector. We show that, in this setting, the LTL problem can be reformulated as a Least Squares (LS) problem and we exploit a novel meta- algorithm to efficiently solve it. At each iteration the meta-algorithm processes only one dataset. Specifically, it firstly estimates the stochastic LS objective function, by splitting this dataset into two subsets used to train and test the inner algorithm, respectively. Secondly, it performs a stochastic gradient step with the estimated value. Under specific assumptions, we present a bound for the generalization error of our meta-algorithm, which suggests the right splitting parameter to choose. When the hyper-parameters of the problem are fixed, this bound is consistent as the number of tasks grows, even if the sample size is kept constant. Preliminary experiments confirm our theoretical findings, highlighting the advantage of our approach, with respect to independent task learning.

Hybrid-MST: A Hybrid Active Sampling Strategy for Pairwise Preference Aggregation

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #126

JING LI · Rafal Mantiuk · Junle Wang · Suiyi Ling · Patrick Le Callet

In this paper we present a hybrid active sampling strategy for pairwise preference aggregation, which aims at recovering the underlying rating of the test candidates from sparse and noisy pairwise labeling. Our method employs Bayesian optimization framework and Bradley-Terry model to construct the utility function, then to obtain the Expected Information Gain (EIG) of each pair. For computational efficiency, Gaussian-Hermite quadrature is used for estimation of EIG. In this work, a hybrid active sampling strategy is proposed, either using Global Maximum (GM) EIG sampling or Minimum Spanning Tree (MST) sampling in each trial, which is determined by the test budget. The proposed method has been validated on both simulated and real-world datasets, where it shows higher preference aggregation ability than the state-of-the-art methods.

Algorithmic Assurance: An Active Approach to Algorithmic Testing using Bayesian Optimisation

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #127

Shivapratap Gopakumar · Sunil Gupta · Santu Rana · Vu Nguyen · Svetha Venkatesh

We introduce algorithmic assurance, the problem of testing whether machine learning algorithms are conforming to their intended design goal. We address this problem by proposing an efficient framework for algorithmic testing. To provide assurance, we need to efficiently discover scenarios where an algorithm decision deviates maximally from its intended gold standard. We mathematically formulate this task as an optimisation problem of an expensive, black-box function. We use an active learning approach based on Bayesian optimisation to solve this optimisation problem. We extend this framework to algorithms with vector-valued outputs by making appropriate modification in Bayesian optimisation via the EXP3 algorithm. We theoretically analyse our methods for convergence. Using two real-world applications, we demonstrate the efficiency of our methods. The significance of our problem formulation and initial solutions is that it will serve as the foundation in assuring humans about machines making complex decisions.

Understanding the Role of Adaptivity in Machine Teaching: The Case of Version Space Learners

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #128

Yuxin Chen · Adish Singla · Oisin Mac Aodha · Pietro Perona · Yisong Yue

In real-world applications of education, an effective teacher adaptively chooses the next example to teach based on the learner’s current state. However, most existing work in algorithmic machine teaching focuses on the batch setting, where adaptivity plays no role. In this paper, we study the case of teaching consistent, version space learners in an interactive setting. At any time step, the teacher provides an example, the learner performs an update, and the teacher observes the learner’s new state. We highlight that adaptivity does not speed up the teaching process when considering existing models of version space learners, such as the “worst-case” model (the learner picks the next hypothesis randomly from the version space) and the “preference-based” model (the learner picks hypothesis according to some global preference). Inspired by human teaching, we propose a new model where the learner picks hypotheses according to some local preference defined by the current hypothesis. We show that our model exhibits several desirable properties, e.g., adaptivity plays a key role, and the learner’s transitions over hypotheses are smooth/interpretable. We develop adaptive teaching algorithms, and demonstrate our results via simulation and user studies.

Active Learning for Non-Parametric Regression Using Purely Random Trees

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #129

Jack Goetz · Ambuj Tewari · Paul Zimmerman

Active learning is the task of using labelled data to select additional points to label, with the goal of fitting the most accurate model with a fixed budget of labelled points. In binary classification active learning is known to produce faster rates than passive learning for a broad range of settings. However in regression restrictive structure and tailored methods were previously needed to obtain theoretically superior performance. In this paper we propose an intuitive tree based active learning algorithm for non-parametric regression with provable improvement over random sampling. When implemented with Mondrian Trees our algorithm is tuning parameter free, consistent and minimax optimal for Lipschitz functions.

Interactive Structure Learning with Structural Query-by-Committee

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #130

Christopher Tosh · Sanjoy Dasgupta

In this work, we introduce interactive structure learning, a framework that unifies many different interactive learning tasks. We present a generalization of the query-by-committee active learning algorithm for this setting, and we study its consistency and rate of convergence, both theoretically and empirically, with and without noise.

Efficient nonmyopic batch active search

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #131

Shali Jiang · Gustavo Malkomes · Matthew Abbott · Benjamin Moseley · Roman Garnett

Active search is a learning paradigm for actively identifying as many members of a given class as possible. A critical target scenario is high-throughput screening for scientific discovery, such as drug or materials discovery. In these settings, specialized instruments can often evaluate \emph{multiple} points simultaneously; however, all existing work on active search focuses on sequential acquisition. We bridge this gap, addressing batch active search from both the theoretical and practical perspective. We first derive the Bayesian optimal policy for this problem, then prove a lower bound on the performance gap between sequential and batch optimal policies: the ``cost of parallelization.'' We also propose novel, efficient batch policies inspired by state-of-the-art sequential policies, and develop an aggressive pruning technique that can dramatically speed up computation. We conduct thorough experiments on data from three application domains: a citation network, material science, and drug discovery, testing all proposed policies (14 total) with a wide range of batch sizes. Our results demonstrate that the empirical performance gap matches our theoretical bound, that nonmyopic policies usually significantly outperform myopic alternatives, and that diversity is an important consideration for batch policy design.

Uncertainty Sampling is Preconditioned Stochastic Gradient Descent on Zero-One Loss

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #132

Stephen Mussmann · Percy Liang

Uncertainty sampling, a popular active learning algorithm, is used to reduce the amount of data required to learn a classifier, but it has been observed in practice to converge to different parameters depending on the initialization and sometimes to even better parameters than standard training on all the data. In this work, we give a theoretical explanation of this phenomenon, showing that uncertainty sampling on a convex (e.g., logistic) loss can be interpreted as performing a preconditioned stochastic gradient step on the population zero-one loss. Experiments on synthetic and real datasets support this connection.

Online Adaptive Methods, Universality and Acceleration

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #133

Yehuda Kfir Levy · Alp Yurtsever · Volkan Cevher

We present a novel method for convex unconstrained optimization that, without any modifications ensures: (1) accelerated convergence rate for smooth objectives, (2) standard convergence rate in the general (non-smooth) setting, and (3) standard convergence rate in the stochastic optimization setting. To the best of our knowledge, this is the first method that simultaneously applies to all of the above settings. At the heart of our method is an adaptive learning rate rule that employs importance weights, in the spirit of adaptive online learning algorithms [duchi2011adaptive,levy2017online], combined with an update that linearly couples two sequences, in the spirit of [AllenOrecchia2017]. An empirical examination of our method demonstrates its applicability to the above mentioned scenarios and corroborates our theoretical findings.

Online Improper Learning with an Approximation Oracle

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #134

Elad Hazan · Wei Hu · Yuanzhi Li · zhiyuan li

We study the following question: given an efficient approximation algorithm for an optimization problem, can we learn efficiently in the same setting? We give a formal affirmative answer to this question in the form of a reduction from online learning to offline approximate optimization using an efficient algorithm that guarantees near optimal regret. The algorithm is efficient in terms of the number of oracle calls to a given approximation oracle – it makes only logarithmically many such calls per iteration. This resolves an open question by Kalai and Vempala, and by Garber. Furthermore, our result applies to the more general improper learning problems.

Online Structured Laplace Approximations for Overcoming Catastrophic Forgetting

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #135

Hippolyt Ritter · Aleksandar Botev · David Barber

We introduce the Kronecker factored online Laplace approximation for overcoming catastrophic forgetting in neural networks. The method is grounded in a Bayesian online learning framework, where we recursively approximate the posterior after every task with a Gaussian, leading to a quadratic penalty on changes to the weights. The Laplace approximation requires calculating the Hessian around a mode, which is typically intractable for modern architectures. In order to make our method scalable, we leverage recent block-diagonal Kronecker factored approximations to the curvature. Our algorithm achieves over 90% test accuracy across a sequence of 50 instantiations of the permuted MNIST dataset, substantially outperforming related methods for overcoming catastrophic forgetting.

Approximating Real-Time Recurrent Learning with Random Kronecker Factors

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #136

Asier Mujika · Florian Meier · Angelika Steger

Despite all the impressive advances of recurrent neural networks, sequential data is still in need of better modelling. Truncated backpropagation through time (TBPTT), the learning algorithm most widely used in practice, suffers from the truncation bias, which drastically limits its ability to learn long-term dependencies.The Real Time Recurrent Learning algorithm (RTRL) addresses this issue, but its high computational requirements make it infeasible in practice. The Unbiased Online Recurrent Optimization algorithm (UORO) approximates RTRL with a smaller runtime and memory cost, but with the disadvantage of obtaining noisy gradients that also limit its practical applicability. In this paper we propose the Kronecker Factored RTRL (KF-RTRL) algorithm that uses a Kronecker product decomposition to approximate the gradients for a large class of RNNs. We show that KF-RTRL is an unbiased and memory efficient online learning algorithm. Our theoretical analysis shows that, under reasonable assumptions, the noise introduced by our algorithm is not only stable over time but also asymptotically much smaller than the one of the UORO algorithm. We also confirm these theoretical results experimentally. Further, we show empirically that the KF-RTRL algorithm captures long-term dependencies and almost matches the performance of TBPTT on real world tasks by training Recurrent Highway Networks on a synthetic string memorization task and on the Penn TreeBank task, respectively. These results indicate that RTRL based approaches might be a promising future alternative to TBPTT.

Online Reciprocal Recommendation with Theoretical Performance Guarantees

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #137

Claudio Gentile · Nikos Parotsidis · Fabio Vitale

A reciprocal recommendation problem is one where the goal of learning is not just to predict a user's preference towards a passive item (e.g., a book), but to recommend the targeted user on one side another user from the other side such that a mutual interest between the two exists. The problem thus is sharply different from the more traditional items-to-users recommendation, since a good match requires meeting the preferences of both users. We initiate a rigorous theoretical investigation of the reciprocal recommendation task in a specific framework of sequential learning. We point out general limitations, formulate reasonable assumptions enabling effective learning and, under these assumptions, we design and analyze a computationally efficient algorithm that uncovers mutual likes at a pace comparable to those achieved by a clairvoyant algorithm knowing all user preferences in advance. Finally, we validate our algorithm against synthetic and real-world datasets, showing improved empirical performance over simple baselines.

Generalized Inverse Optimization through Online Learning

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #138

Chaosheng Dong · Yiran Chen · Bo Zeng

Inverse optimization is a powerful paradigm for learning preferences and restrictions that explain the behavior of a decision maker, based on a set of external signal and the corresponding decision pairs. However, most inverse optimization algorithms are designed specifically in batch setting, where all the data is available in advance. As a consequence, there has been rare use of these methods in an online setting suitable for real-time applications. In this paper, we propose a general framework for inverse optimization through online learning. Specifically, we develop an online learning algorithm that uses an implicit update rule which can handle noisy data. Moreover, under additional regularity assumptions in terms of the data and the model, we prove that our algorithm converges at a rate of $\mathcal{O}(1/\sqrt{T})$ and is statistically consistent. In our experiments, we show the online learning approach can learn the parameters with great accuracy and is very robust to noises, and achieves a dramatic improvement in computational efficacy over the batch learning approach.

Adaptive Online Learning in Dynamic Environments

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #139

Lijun Zhang · Shiyin Lu · Zhi-Hua Zhou

In this paper, we study online convex optimization in dynamic environments, and aim to bound the dynamic regret with respect to any sequence of comparators. Existing work have shown that online gradient descent enjoys an $O(\sqrt{T}(1+P_T))$ dynamic regret, where T is the number of iterations and P_T is the path-length of the comparator sequence. However, this result is unsatisfactory, as there exists a large gap from the $\Omega(\sqrt{T(1+P_T)})$ lower bound established in our paper. To address this limitation, we develop a novel online method, namely adaptive learning for dynamic environment (Ader), which achieves an optimal $O(\sqrt{T(1+P_T)})$ dynamic regret. The basic idea is to maintain a set of experts, each attaining an optimal dynamic regret for a specific path-length, and combines them with an expert-tracking algorithm. Furthermore, we propose an improved Ader based on the surrogate loss, and in this way the number of gradient evaluations per round is reduced from $O(\log T)$ to 1. Finally, we extend Ader to the setting that a sequence of dynamical models is available to characterize the comparators.

Online convex optimization for cumulative constraints

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #140

Jianjun Yuan · Andrew Lamperski

We propose the algorithms for online convex optimization which lead to cumulative squared constraint violations of the form $\sum\limits_{t=1}^T\big([g(x_t)]+*big)^2=O(T^{1-\beta})$, where $\beta\in(0,1)$. Previous literature has focused on long-term constraints of the form $\sum\limits*{t=1}^Tg(x_t)$. There, strictly feasible solutions can cancel out the effects of violated constraints. In contrast, the new form heavily penalizes large constraint violations and cancellation effects cannot occur. Furthermore, useful bounds on the single step constraint violation $[g(x_t)]_+$ are derived. For convex objectives, our regret bounds generalize existing bounds, and for strongly convex objectives we give improved regret bounds. In numerical experiments, we show that our algorithm closely follows the constraint boundary leading to low cumulative violation.

Efficient online algorithms for fast-rate regret bounds under sparsity

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #141

Pierre Gaillard · Olivier Wintenberger

We consider the problem of online convex optimization in two different settings: arbitrary and i.i.d. sequence of convex loss functions. In both settings, we provide efficient algorithms whose cumulative excess risks are controlled with fast-rate sparse bounds. First, the excess risks bounds depend on the sparsity of the objective rather than on the dimension of the parameters space. Second, their rates are faster than the slow-rate $1/\sqrt{T}$ under additional convexity assumptions on the loss functions. In the adversarial setting, we develop an algorithm BOA+ whose cumulative excess risks is controlled by several bounds with different trade-offs between sparsity and rate for strongly convex loss functions. In the i.i.d. setting under the Łojasiewicz's assumption, we establish new risk bounds that are sparse with a rate adaptive to the convexity of the risk (ranging from a rate $1/\sqrt{T}$ for general convex risk to $1/T$ for strongly convex risk). These results generalize previous works on sparse online learning under weak assumptions on the risk.

Regret Bounds for Online Portfolio Selection with a Cardinality Constraint

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #142

Shinji Ito · Daisuke Hatano · Sumita Hanna · Akihiro Yabe · Takuro Fukunaga · Naonori Kakimura · Ken-Ichi Kawarabayashi

Online portfolio selection is a sequential decision-making problem in which a learner repetitively selects a portfolio over a set of assets, aiming to maximize long-term return. In this paper, we study the problem with the cardinality constraint that the number of assets in a portfolio is restricted to be at most k, and consider two scenarios: (i) in the full-feedback setting, the learner can observe price relatives (rates of return to cost) for all assets, and (ii) in the bandit-feedback setting, the learner can observe price relatives only for invested assets. We propose efficient algorithms for these scenarios that achieve sublinear regrets. We also provide regret (statistical) lower bounds for both scenarios which nearly match the upper bounds when k is a constant. In addition, we give a computational lower bound which implies that no algorithm maintains both computational efficiency, as well as a small regret upper bound.

Policy Regret in Repeated Games

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #144

Raman Arora · Michael Dinitz · Teodor Vanislavov Marinov · Mehryar Mohri

The notion of policy regret'' in online learning is supposed to capture the reactions of the adversary to the actions taken by the learner, which more traditional notions such as external regret do not take into account. We revisit this notion of policy regret, and first show that there are online learning settings in which policy regret and external regret are incompatible: any sequence of play which does well with respect to one must do poorly with respect to the other. We then focus on the game theoretic setting, when the adversary is a self-interested agent. In this setting we show that the external regret and policy regret are not in conflict, and in fact that a wide class of algorithms can ensure both as long as the adversary is also using such an algorithm. We also define a new notion of equilibrium which we call apolicy equilibrium'', and show that no-policy regret algorithms will have play which converges to such an equilibrium. Relating this back to external regret, we show that coarse correlated equilibria (which no-external regret players will converge to) are a strict subset of policy equilibria. So in game-theoretic settings every sequence of play with no external regret also has no policy regret, but the converse is not true.

Query Complexity of Bayesian Private Learning

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #145

Kuang Xu

We study the query complexity of Bayesian Private Learning: a learner wishes to locate a random target within an interval by submitting queries, in the presence of an adversary who observes all of her queries but not the responses. How many queries are necessary and sufficient in order for the learner to accurately estimate the target, while simultaneously concealing the target from the adversary?

Our main result is a query complexity lower bound that is tight up to the first order. We show that if the learner wants to estimate the target within an error of ϵ, while ensuring that no adversary estimator can achieve a constant additive error with probability greater than $1/L$, then the query complexity is on the order of $L\log(1/\epsilon)$ as $\epsilon \to 0$. Our result demonstrates that increased privacy, as captured by L, comes at the expense of a \emph{multiplicative} increase in query complexity. The proof builds on Fano's inequality and properties of certain proportional-sampling estimators.

The Limits of Post-Selection Generalization

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #146

Jonathan Ullman · Adam Smith · Kobbi Nissim · Uri Stemmer · Thomas Steinke

While statistics and machine learning offers numerous methods for ensuring generalization, these methods often fail in the presence of post selection---the common practice in which the choice of analysis depends on previous interactions with the same dataset. A recent line of work has introduced powerful, general purpose algorithms that ensure a property called post hoc generalization (Cummings et al., COLT'16), which says that no person when given the output of the algorithm should be able to find any statistic for which the data differs significantly from the population it came from. In this work we show several limitations on the power of algorithms satisfying post hoc generalization. First, we show a tight lower bound on the error of any algorithm that satisfies post hoc generalization and answers adaptively chosen statistical queries, showing a strong barrier to progress in post selection data analysis. Second, we show that post hoc generalization is not closed under composition, despite many examples of such algorithms exhibiting strong composition properties.

Sequential Test for the Lowest Mean: From Thompson to Murphy Sampling

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #147

Emilie Kaufmann · Wouter Koolen · Aurélien Garivier

Learning the minimum/maximum mean among a finite set of distributions is a fundamental sub-problem in planning, game tree search and reinforcement learning. We formalize this learning task as the problem of sequentially testing how the minimum mean among a finite set of distributions compares to a given threshold. We develop refined non-asymptotic lower bounds, which show that optimality mandates very different sampling behavior for a low vs high true minimum. We show that Thompson Sampling and the intuitive Lower Confidence Bounds policy each nail only one of these cases. We develop a novel approach that we call Murphy Sampling. Even though it entertains exclusively low true minima, we prove that MS is optimal for both possibilities. We then design advanced self-normalized deviation inequalities, fueling more aggressive stopping rules. We complement our theoretical guarantees by experiments showing that MS works best in practice.

Contextual Combinatorial Multi-armed Bandits with Volatile Arms and Submodular Reward

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #148

Lixing Chen · Jie Xu · Zhuo Lu

In this paper, we study the stochastic contextual combinatorial multi-armed bandit (CC-MAB) framework that is tailored for volatile arms and submodular reward functions. CC-MAB inherits properties from both contextual bandit and combinatorial bandit: it aims to select a set of arms in each round based on the side information (a.k.a. context) associated with the arms. By `volatile arms'', we mean that the available arms to select from in each round may change; and by`submodular rewards'', we mean that the total reward achieved by selected arms is not a simple sum of individual rewards but demonstrates a feature of diminishing returns determined by the relations between selected arms (e.g. relevance and redundancy). Volatile arms and submodular rewards are often seen in many real-world applications, e.g. recommender systems and crowdsourcing, in which multi-armed bandit (MAB) based strategies are extensively applied. Although there exist works that investigate these issues separately based on standard MAB, jointly considering all these issues in a single MAB problem requires very different algorithm design and regret analysis. Our algorithm CC-MAB provides an online decision-making policy in a contextual and combinatorial bandit setting and effectively addresses the issues raised by volatile arms and submodular reward functions. The proposed algorithm is proved to achieve $O(cT^{\frac{2\alpha+D}{3\alpha + D}}\log(T))$ regret after a span of T rounds. The performance of CC-MAB is evaluated by experiments conducted on a real-world crowdsourcing dataset, and the result shows that our algorithm outperforms the prior art.

TopRank: A practical algorithm for online stochastic ranking

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #149

Tor Lattimore · Branislav Kveton · Shuai Li · Csaba Szepesvari

Online learning to rank is a sequential decision-making problem where in each round the learning agent chooses a list of items and receives feedback in the form of clicks from the user. Many sample-efficient algorithms have been proposed for this problem that assume a specific click model connecting rankings and user behavior. We propose a generalized click model that encompasses many existing models, including the position-based and cascade models. Our generalization motivates a novel online learning algorithm based on topological sort, which we call TopRank. TopRank is (a) more natural than existing algorithms, (b) has stronger regret guarantees than existing algorithms with comparable generality, (c) has a more insightful proof that leaves the door open to many generalizations, (d) outperforms existing algorithms empirically.

A Bandit Approach to Sequential Experimental Design with False Discovery Control

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #150

Kevin Jamieson · Lalit Jain

We propose a new adaptive sampling approach to multiple testing which aims to maximize statistical power while ensuring anytime false discovery control. We consider n distributions whose means are partitioned by whether they are below or equal to a baseline (nulls), versus above the baseline (true positives). In addition, each distribution can be sequentially and repeatedly sampled. Using techniques from multi-armed bandits, we provide an algorithm that takes as few samples as possible to exceed a target true positive proportion (i.e. proportion of true positives discovered) while giving anytime control of the false discovery proportion (nulls predicted as true positives). Our sample complexity results match known information theoretic lower bounds and through simulations we show a substantial performance improvement over uniform sampling and an adaptive elimination style algorithm. Given the simplicity of the approach, and its sample efficiency, the method has promise for wide adoption in the biological sciences, clinical testing for drug discovery, and maximization of click through in A/B/n testing problems.

Adaptation to Easy Data in Prediction with Limited Advice

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #151

Tobias Thune · Yevgeny Seldin

We derive an online learning algorithm with improved regret guarantees for `easy'' loss sequences. We consider two types of`easiness'': (a) stochastic loss sequences and (b) adversarial loss sequences with small effective range of the losses. While a number of algorithms have been proposed for exploiting small effective range in the full information setting, Gerchinovitz and Lattimore [2016] have shown the impossibility of regret scaling with the effective range of the losses in the bandit setting. We show that just one additional observation per round is sufficient to circumvent the impossibility result. The proposed Second Order Difference Adjustments (SODA) algorithm requires no prior knowledge of the effective range of the losses, ε, and achieves an $O(\varepsilon \sqrt{KT \ln K}) + \tilde{O}(\varepsilon K \sqrt[4]{T})$ expected regret guarantee, where T is the time horizon and K is the number of actions. The scaling with the effective loss range is achieved under significantly weaker assumptions than those made by Cesa-Bianchi and Shamir [2018] in an earlier attempt to circumvent the impossibility result. We also provide a regret lower bound of $\Omega(\varepsilon\sqrt{T K})$, which almost matches the upper bound. In addition, we show that in the stochastic setting SODA achieves an $O\left(\sum_{a:\Delta_a>0} \frac{K\varepsilon^2}{\Delta_a}\right)$ pseudo-regret bound that holds simultaneously with the adversarial regret guarantee. In other words, SODA is safe against an unrestricted oblivious adversary and provides improved regret guarantees for at least two different types of ``easiness'' simultaneously.

Differentially Private Contextual Linear Bandits

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #152

Roshan Shariff · Or Sheffet

We study the contextual linear bandit problem, a version of the standard stochastic multi-armed bandit (MAB) problem where a learner sequentially selects actions to maximize a reward which depends also on a user provided per-round context. Though the context is chosen arbitrarily or adversarially, the reward is assumed to be a stochastic function of a feature vector that encodes the context and selected action. Our goal is to devise private learners for the contextual linear bandit problem. We first show that using the standard definition of differential privacy results in linear regret. So instead, we adopt the notion of joint differential privacy, where we assume that the action chosen on day t is only revealed to user t and thus needn't be kept private that day, only on following days. We give a general scheme converting the classic linear-UCB algorithm into a joint differentially private algorithm using the tree-based algorithm. We then apply either Gaussian noise or Wishart noise to achieve joint-differentially private algorithms and bound the resulting algorithms' regrets. In addition, we give the first lower bound on the additional regret any private algorithms for the MAB problem must incur.

Community Exploration: From Offline Optimization to Online Learning

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #153

Xiaowei Chen · Weiran Huang · Wei Chen · John C. S. Lui

We introduce the community exploration problem that has various real-world applications such as online advertising. In the problem, an explorer allocates limited budget to explore communities so as to maximize the number of members he could meet. We provide a systematic study of the community exploration problem, from offline optimization to online learning. For the offline setting where the sizes of communities are known, we prove that the greedy methods for both of non-adaptive exploration and adaptive exploration are optimal. For the online setting where the sizes of communities are not known and need to be learned from the multi-round explorations, we propose an ``upper confidence'' like algorithm that achieves the logarithmic regret bounds. By combining the feedback from different rounds, we can achieve a constant regret bound.

Adaptive Learning with Unknown Information Flows

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #154

Yonatan Gur · Ahmadreza Momeni

An agent facing sequential decisions that are characterized by partial feedback needs to strike a balance between maximizing immediate payoffs based on available information, and acquiring new information that may be essential for maximizing future payoffs. This trade-off is captured by the multi-armed bandit (MAB) framework that has been studied and applied when at each time epoch payoff observations are collected on the actions that are selected at that epoch. In this paper we introduce a new, generalized MAB formulation in which additional information on each arm may appear arbitrarily throughout the decision horizon, and study the impact of such information flows on the achievable performance and the design of efficient decision-making policies. By obtaining matching lower and upper bounds, we characterize the (regret) complexity of this family of MAB problems as a function of the information flows. We introduce an adaptive exploration policy that, without any prior knowledge of the information arrival process, attains the best performance (in terms of regret rate) that is achievable when the information arrival process is a priori known. Our policy uses dynamically customized virtual time indexes to endogenously control the exploration rate based on the realized information arrival process.

Multi-armed Bandits with Compensation

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #155

Siwei Wang · Longbo Huang

We propose and study the known-compensation multi-arm bandit (KCMAB) problem, where a system controller offers a set of arms to many short-term players for T steps. In each step, one short-term player arrives to the system. Upon arrival, the player greedily selects an arm with the current best average reward and receives a stochastic reward associated with the arm. In order to incentivize players to explore other arms, the controller provides proper payment compensation to players. The objective of the controller is to maximize the total reward collected by players while minimizing the compensation. We first give a compensation lower bound $\Theta(\sum_i {\Delta_i\log T\over KL_i})$, where Δ_i and KL_i are the expected reward gap and Kullback-Leibler (KL) divergence between distributions of arm i and the best arm, respectively. We then analyze three algorithms to solve the KCMAB problem, and obtain their regrets and compensations. We show that the algorithms all achieve $O(\log T)$ regret and $O(\log T)$ compensation that match the theoretical lower bound. Finally, we use experiments to show the behaviors of those algorithms.

Bandit Learning with Implicit Feedback

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #156

Yi Qi · Qingyun Wu · Hongning Wang · Jie Tang · Maosong Sun

Implicit feedback, such as user clicks, although abundant in online information service systems, does not provide substantial evidence on users' evaluation of system's output. Without proper modeling, such incomplete supervision inevitably misleads model estimation, especially in a bandit learning setting where the feedback is acquired on the fly. In this work, we perform contextual bandit learning with implicit feedback by modeling the feedback as a composition of user result examination and relevance judgment. Since users' examination behavior is unobserved, we introduce latent variables to model it. We perform Thompson sampling on top of variational Bayesian inference for arm selection and model update. Our upper regret bound analysis of the proposed algorithm proves its feasibility of learning from implicit feedback in a bandit setting; and extensive empirical evaluations on click logs collected from a major MOOC platform further demonstrate its learning effectiveness in practice.

Optimistic optimization of a Brownian

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #157

Jean-Bastien Grill · Michal Valko · Remi Munos

We address the problem of optimizing a Brownian motion. We consider a (random) realization W of a Brownian motion with input space in $[0,1]$. Given W, our goal is to return an ϵ-approximation of its maximum using the smallest possible number of function evaluations, the sample complexity of the algorithm. We provide an algorithm with sample complexity of order $\log^2(1/\epsilon)$. This improves over previous results of Al-Mharmah and Calvin (1996) and Calvin et al. (2017) which provided only polynomial rates. Our algorithm is adaptive---each query depends on previous values---and is an instance of the optimism-in-the-face-of-uncertainty principle.

Bandit Learning with Positive Externalities

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #158

Virag Shah · Jose Blanchet · Ramesh Johari

In many platforms, user arrivals exhibit a self-reinforcing behavior: future user arrivals are likely to have preferences similar to users who were satisfied in the past. In other words, arrivals exhibit {\em positive externalities}. We study multiarmed bandit (MAB) problems with positive externalities. We show that the self-reinforcing preferences may lead standard benchmark algorithms such as UCB to exhibit linear regret. We develop a new algorithm, Balanced Exploration (BE), which explores arms carefully to avoid suboptimal convergence of arrivals before sufficient evidence is gathered. We also introduce an adaptive variant of BE which successively eliminates suboptimal arms. We analyze their asymptotic regret, and establish optimality by showing that no algorithm can perform better.

An Information-Theoretic Analysis for Thompson Sampling with Many Actions

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #159

Shi Dong · Benjamin Van Roy

Information-theoretic Bayesian regret bounds of Russo and Van Roy capture the dependence of regret on prior uncertainty. However, this dependence is through entropy, which can become arbitrarily large as the number of actions increases. We establish new bounds that depend instead on a notion of rate-distortion. Among other things, this allows us to recover through information-theoretic arguments a near-optimal bound for the linear bandit. We also offer a bound for the logistic bandit that dramatically improves on the best previously available, though this bound depends on an information-theoretic statistic that we have only been able to quantify via computation.

Distributed Multi-Player Bandits - a Game of Thrones Approach

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #160

Ilai Bistritz · Amir Leshem

We consider a multi-armed bandit game where N players compete for K arms for T turns. Each player has different expected rewards for the arms, and the instantaneous rewards are independent and identically distributed. Performance is measured using the expected sum of regrets, compared to the optimal assignment of arms to players. We assume that each player only knows her actions and the reward she received each turn. Players cannot observe the actions of other players, and no communication between players is possible. We present a distributed algorithm and prove that it achieves an expected sum of regrets of near-O\left(\log^{2}T\right). This is the first algorithm to achieve a poly-logarithmic regret in this fully distributed scenario. All other works have assumed that either all players have the same vector of expected rewards or that communication between players is possible.

PG-TS: Improved Thompson Sampling for Logistic Contextual Bandits

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #161

Bianca Dumitrascu · Karen Feng · Barbara Engelhardt

We address the problem of regret minimization in logistic contextual bandits, where a learner decides among sequential actions or arms given their respective contexts to maximize binary rewards. Using a fast inference procedure with Polya-Gamma distributed augmentation variables, we propose an improved version of Thompson Sampling, a Bayesian formulation of contextual bandits with near-optimal performance. Our approach, Polya-Gamma augmented Thompson Sampling (PG-TS), achieves state-of-the-art performance on simulated and real data. PG-TS explores the action space efficiently and exploits high-reward arms, quickly converging to solutions of low regret. Its explicit estimation of the posterior distribution of the context feature covariance leads to substantial empirical gains over approximate approaches. PG-TS is the first approach to demonstrate the benefits of Polya-Gamma augmentation in bandits and to propose an efficient Gibbs sampler for approximating the analytically unsolvable integral of logistic contextual bandits.

Non-delusional Q-learning and value-iteration

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #162

Tyler Lu · Dale Schuurmans · Craig Boutilier

We identify a fundamental source of error in Q-learning and other forms of dynamic programming with function approximation. Delusional bias arises when the approximation architecture limits the class of expressible greedy policies. Since standard Q-updates make globally uncoordinated action choices with respect to the expressible policy class, inconsistent or even conflicting Q-value estimates can result, leading to pathological behaviour such as over/under-estimation, instability and even divergence. To solve this problem, we introduce a new notion of policy consistency and define a local backup process that ensures global consistency through the use of information sets---sets that record constraints on policies consistent with backed-up Q-values. We prove that both the model-based and model-free algorithms using this backup remove delusional bias, yielding the first known algorithms that guarantee optimal results under general conditions. These algorithms furthermore only require polynomially many information sets (from a potentially exponential support). Finally, we suggest other practical heuristics for value-iteration and Q-learning that attempt to reduce delusional bias.

Differentiable MPC for End-to-end Planning and Control

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #163

Brandon Amos · Ivan Jimenez · Jacob Sacks · Byron Boots · J. Zico Kolter

We present foundations for using Model Predictive Control (MPC) as a differentiable policy class for reinforcement learning. This provides one way of leveraging and combining the advantages of model-free and model-based approaches. Specifically, we differentiate through MPC by using the KKT conditions of the convex approximation at a fixed point of the controller. Using this strategy, we are able to learn the cost and dynamics of a controller via end-to-end learning. Our experiments focus on imitation learning in the pendulum and cartpole domains, where we learn the cost and dynamics terms of an MPC policy class. We show that our MPC policies are significantly more data-efficient than a generic neural network and that our method is superior to traditional system identification in a setting where the expert is unrealizable.

Multiple-Step Greedy Policies in Approximate and Online Reinforcement Learning

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #164

Yonathan Efroni · Gal Dalal · Bruno Scherrer · Shie Mannor

Multiple-step lookahead policies have demonstrated high empirical competence in Reinforcement Learning, via the use of Monte Carlo Tree Search or Model Predictive Control. In a recent work (Efroni et al., 2018), multiple-step greedy policies and their use in vanilla Policy Iteration algorithms were proposed and analyzed. In this work, we study multiple-step greedy algorithms in more practical setups. We begin by highlighting a counter-intuitive difficulty, arising with soft-policy updates: even in the absence of approximations, and contrary to the 1-step-greedy case, monotonic policy improvement is not guaranteed unless the update stepsize is sufficiently large. Taking particular care about this difficulty, we formulate and analyze online and approximate algorithms that use such a multi-step greedy operator.

Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #165

Kurtland Chua · Roberto Calandra · Rowan McAllister · Sergey Levine

Model-based reinforcement learning (RL) algorithms can attain excellent sample efficiency, but often lag behind the best model-free algorithms in terms of asymptotic performance. This is especially true with high-capacity parametric function approximators, such as deep networks. In this paper, we study how to bridge this gap, by employing uncertainty-aware dynamics models. We propose a new algorithm called probabilistic ensembles with trajectory sampling (PETS) that combines uncertainty-aware deep network dynamics models with sampling-based uncertainty propagation. Our comparison to state-of-the-art model-based and model-free deep RL algorithms shows that our approach matches the asymptotic performance of model-free algorithms on several challenging benchmark tasks, while requiring significantly fewer samples (e.g. 8 and 125 times fewer samples than Soft Actor Critic and Proximal Policy Optimization respectively on the half-cheetah task).

Learning convex bounds for linear quadratic control policy synthesis

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #166

Jack Umenberger · Thomas Schön

Learning to make decisions from observed data in dynamic environments remains a problem of fundamental importance in a numbers of fields, from artificial intelligence and robotics, to medicine and finance. This paper concerns the problem of learning control policies for unknown linear dynamical systems so as to maximize a quadratic reward function. We present a method to optimize the expected value of the reward over the posterior distribution of the unknown system parameters, given data. The algorithm involves sequential convex programing, and enjoys reliable local convergence and robust stability guarantees. Numerical simulations and stabilization of a real-world inverted pendulum are used to demonstrate the approach, with strong performance and robustness properties observed in both.

Sample-Efficient Reinforcement Learning with Stochastic Ensemble Value Expansion

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #167

Jacob Buckman · Danijar Hafner · George Tucker · Eugene Brevdo · Honglak Lee

There is growing interest in combining model-free and model-based approaches in reinforcement learning with the goal of achieving the high performance of model-free algorithms with low sample complexity. This is difficult because an imperfect dynamics model can degrade the performance of the learning algorithm, and in sufficiently complex environments, the dynamics model will always be imperfect. As a result, a key challenge is to combine model-based approaches with model-free learning in such a way that errors in the model do not degrade performance. We propose stochastic ensemble value expansion (STEVE), a novel model-based technique that addresses this issue. By dynamically interpolating between model rollouts of various horizon lengths, STEVE ensures that the model is only utilized when doing so does not introduce significant errors. Our approach outperforms model-free baselines on challenging continuous control benchmarks with an order-of-magnitude increase in sample efficiency.

Policy-Conditioned Uncertainty Sets for Robust Markov Decision Processes

Poster | Thu Dec 6th 05:00 -- 07:00 PM @ Room 210 & 230 AB #168

Andrea Tirinzoni · Marek Petrik · Xiangli Chen · Brian Ziebart

What policy should be employed in a Markov decision process with uncertain parameters? Robust optimization answer to this question is to use rectangular uncertainty sets, which independently reflect available knowledge about each state, and then obtains a decision policy that maximizes expected reward for the worst-case decision process parameters from these uncertainty sets. While this rectangularity is convenient computationally and leads to tractable solutions, it often produces policies that are too conservative in practice, and does not facilitate knowledge transfer between portions of the state space or across related decision processes. In this work, we propose non-rectangular uncertainty sets that bound marginal moments of state-action features defined over entire trajectories through a decision process. This enables generalization to different portions of the state space while retaining appropriate uncertainty of the decision process. We develop algorithms for solving the resulting robust decision problems, which reduce to finding an optimal policy for a mixture of decision processes, and demonstrate the benefits of our approach experimentally.