Sign in to GitHub · GitHub

****

Uncovering Causality from Multivariate Hawkes Integrated Cumulants

Massil Achab, Emmanuel Bacry, Stéphane Ga■■ffas, Iacopo Mastromatteo, Jean-François Muzy

We design a new nonparametric method that allows one to estimate the matrix of integrated kernels of a multivariate Hawkes process. This matrix not only encodes the mutual influences of each node of the process, but also disentangles the causality relationships between them. Our approach is the first that leads to an estimation of this matrix without any parametric modeling and estimation of the kernels themselves. A consequence is that it can give an estimation of causality relationships between nodes (or users), based on their activity timestamps (on a social network for instance), without knowing or estimating the shape of the activities lifetime. For that purpose, we introduce a moment matching method that fits the second-order and the third-order integrated cumulants of the process. A theoretical analysis allows to prove that this new estimation technique is consistent. Moreover, we show on numerical experiments that our approach is indeed very robust to the shape of the kernels, and gives appealing results on the MemeTracker database and on financial order book data.

****

A Unified Maximum Likelihood Approach for Estimating Symmetric Properties of Discrete Distributions

Jayadev Acharya, Hirakendu Das, Alon Orlitsky, Ananda Theertha Suresh

Symmetric distribution properties such as support size, support coverage, entropy, and proximity to uniformity, arise in many applications. Recently, researchers applied different estimators and analysis tools to derive asymptotically sample-optimal approximations for each of these properties. We show that a single, simple, plug-in estimator—profile maximum likelihood (PML)—is sample competitive for all symmetric properties, and in particular is asymptotically sample-optimal for all the above properties.

****

Constrained Policy Optimization

Joshua Achiam, David Held, Aviv Tamar, Pieter Abbeel

For many applications of reinforcement learning it can be more convenient to specify both a reward function and constraints, rather than trying to design behavior through the reward function. For example, systems that physically interact with or around humans should satisfy safety constraints. Recent advances in policy search algorithms (Mnih et al., 2016, Schulman et al., 2015, Lillicrap et al., 2016, Levine et al., 2016) have enabled new capabilities in high-dimensional control, but do not consider the constrained setting. We propose Constrained Policy Optimization (CPO), the first general-purpose policy search algorithm for constrained reinforcement learning with guarantees for near-constraint satisfaction at each iteration. Our method allows us to train neural network policies for high-dimensional control while making guarantees about policy behavior all throughout training. Our guarantees are based on a new theoretical result, which is of independent interest: we prove a bound relating the expected returns of two policies to an average divergence between them. We demonstrate the effectiveness of our approach on simulated robot locomotion tasks where the agent must satisfy constraints motivated by safety.

****

The Price of Differential Privacy for Online Learning

Naman Agarwal, Karan Singh

We design differentially private algorithms for the problem of online linear optimization in the full information and bandit settings with optimal $O(T^{0.5})$ regret bounds. In the full-information setting, our results demonstrate that ϵ-differential privacy may be ensured for free – in particular, the regret bounds scale as $O(T^{0.5}+1/\epsilon)$. For bandit linear optimization, and as a special case, for non-stochastic multi-armed bandits, the proposed algorithm achieves a regret of $O(T^{0.5}/\epsilon)$, while the previously best known regret bound was $O(T^{2/3}/\epsilon)$.

****

Local Bayesian Optimization of Motor Skills

Riad Akrour, Dmitry Sorokin, Jan Peters, Gerhard Neumann

Bayesian optimization is renowned for its sample efficiency but its application to higher dimensional tasks is impeded by its focus on global optimization. To scale to higher dimensional problems, we leverage the sample efficiency of Bayesian optimization in a local context. The optimization of the acquisition function is restricted to the vicinity of a Gaussian search distribution which is moved towards high value areas of the objective. The proposed information-theoretic update of the search distribution results in a Bayesian interpretation of local stochastic search: the search distribution encodes prior knowledge on the optimum’s location and is weighted at each iteration by the likelihood of this location’s optimality. We demonstrate the effectiveness of our algorithm on several benchmark objective functions as well as a continuous robotic task in which an informative prior is obtained by imitation learning.

****

Connected Subgraph Detection with Mirror Descent on SDPs

Cem Aksoylar, Lorenzo Orecchia, Venkatesh Saligrama

We propose a novel, computationally efficient mirror-descent based optimization framework for subgraph detection in graph-structured data. Our aim is to discover anomalous patterns present in a connected subgraph of a given graph. This problem arises in many applications such as detection of network intrusions, community detection, detection of anomalous events in surveillance videos or disease outbreaks. Since optimization over connected subgraphs is a combinatorial and computationally difficult problem, we propose a convex relaxation that offers a principled approach to incorporating connectivity and conductance constraints on candidate subgraphs. We develop a novel efficient algorithm to solve the relaxed problem, establish convergence guarantees and demonstrate its feasibility and performance with experiments on real and very large simulated networks.

****

Learning from Clinical Judgments: Semi-Markov-Modulated Marked Hawkes Processes for Risk Prognosis

Ahmed M. Alaa, Scott Hu, Mihaela Schaar

Critically ill patients in regular wards are vulnerable to unanticipated adverse events which require prompt transfer to the intensive care unit (ICU). To allow for accurate prognosis of deteriorating patients, we develop a novel continuous-time probabilistic model for a monitored patient’s temporal sequence of physiological data. Our model captures “informatively sampled” patient episodes: the clinicians’ decisions on when to observe a hospitalized patient’s vital signs and lab tests over time are represented by a marked Hawkes process, with intensity parameters that are modulated by the patient’s latent clinical states, and with observable physiological data (mark process) modeled as a switching multi-task Gaussian process. In addition, our model captures “informatively censored” patient episodes by representing the patient’s latent clinical states as an absorbing semi-Markov jump process. The model parameters are learned from offline patient episodes in the electronic health records via an EM-based algorithm. Experiments conducted on a cohort of patients admitted to a major medical center over a 3-year period show that risk prognosis based on our model significantly outperforms the currently deployed medical risk scores and other baseline machine learning algorithms.

****

A Semismooth Newton Method for Fast, Generic Convex Programming

Alnur Ali, Eric Wong, J. Zico Kolter

We introduce Newton-ADMM, a method for fast conic optimization. The basic idea is to view the residuals of consecutive iterates generated by the alternating direction method of multipliers (ADMM) as a set of fixed point equations, and then use a nonsmooth Newton method to find a solution; we apply the basic idea to the Splitting Cone Solver (SCS), a state-of-the-art method for solving generic conic optimization problems. We demonstrate theoretically, by extending the theory of semismooth operators, that Newton-ADMM converges rapidly (i.e., quadratically) to a solution; empirically, Newton-ADMM is significantly faster than SCS on a number of problems. The method also has essentially no tuning parameters, generates certificates of primal or dual infeasibility, when appropriate, and can be specialized to solve specific convex problems.

****

Learning Continuous Semantic Representations of Symbolic Expressions

Miltiadis Allamanis, Pankajan Chanthirasegaran, Pushmeet Kohli, Charles Sutton

Combining abstract, symbolic reasoning with continuous neural reasoning is a grand challenge of representation learning. As a step in this direction, we propose a new architecture, called neural equivalence network, for the problem of learning continuous semantic representations of algebraic and logical expressions. These networks are trained to represent semantic equivalence, even of expressions that are syntactically very different. The challenge is that semantic representations must be computed in a syntax-directed manner, because semantics is compositional, but at the same time, small changes in syntax can lead to very large changes in semantics, which can be difficult for continuous neural architectures. We perform an exhaustive evaluation on the task of checking equivalence on a highly diverse class of symbolic algebraic and boolean expression types, showing that our model significantly outperforms existing architectures.

****

Natasha: Faster Non-Convex Stochastic Optimization via Strongly Non-Convex Parameter

Zeyuan Allen-Zhu

Given a non-convex function $f(x)$ that is an average of n smooth functions, we design stochastic first-order methods to find its approximate stationary points. The performance of our new methods depend on the smallest (negative) eigenvalue $-\sigma$ of the Hessian. This parameter σ captures how strongly non-convex $f(x)$ is, and is analogous to the strong convexity parameter for convex optimization. At least in theory, our methods outperform known results for a range of parameter σ, and can also be used to find approximate local minima. Our result implies an interesting dichotomy: there exists a threshold σ_0 so that the (currently) fastest methods for $\sigma>\sigma_0$ and for $\sigma<\sigma_0$ have different behaviors: the former scales with $n^{2/3}$ and the latter scales with $n^{3/4}$.

****

Doubly Accelerated Methods for Faster CCA and Generalized Eigendecomposition

Zeyuan Allen-Zhu, Yuanzhi Li

We study k-GenEV, the problem of finding the top k generalized eigenvectors, and k-CCA, the problem of finding the top k vectors in canonical-correlation analysis. We propose algorithms LazyEV and LazyCCA to solve the two problems with running times linearly dependent on the input size and on k. Furthermore, our algorithms are doubly-accelerated: our running times depend only on the square root of the matrix condition number, and on the square root of the eigengap. This is the first such result for both k-GenEV or k-CCA. We also provide the first gap-free results, which provide running times that depend on $1/\sqrt{\varepsilon}$ rather than the eigengap.

****

Faster Principal Component Regression and Stable Matrix Chebyshev Approximation

Zeyuan Allen-Zhu, Yuanzhi Li

We solve principal component regression (PCR), up to a multiplicative accuracy $1+\gamma$, by reducing the problem to $\tilde{O}(\gamma^{-1})$ black-box calls of ridge regression. Therefore, our algorithm does not require any explicit construction of the top principal components, and is suitable for large-scale PCR instances. In contrast, previous result requires $\tilde{O}(\gamma^{-2})$ such black-box calls. We obtain this result by developing a general stable recurrence formula for matrix Chebyshev polynomials, and a degree-optimal polynomial approximation to the matrix sign function. Our techniques may be of independent interests, especially when designing iterative methods.

****

Follow the Compressed Leader: Faster Online Learning of Eigenvectors and Faster

MMWU

Zeyuan Allen-Zhu, Yuanzhi Li

The online problem of computing the top eigenvector is fundamental to machine learning. The famous matrix-multiplicative-weight-update (MMWU) framework solves this online problem and gives optimal regret. However, since MMWU runs very slow due to the computation of matrix exponentials, researchers proposed the follow-the-perturbed-leader (FTPL) framework which is faster, but a factor $\sqrt{d}$ worse than the optimal regret for dimension-d matrices. We propose a follow-the-compressed-leader framework which, not only matches the optimal regret of MMWU (up to polylog factors), but runs no slower than FTPL. Our main idea is to “compress” the MMWU strategy to dimension 3 in the adversarial setting, or dimension 1 in the stochastic setting. This resolves an open question regarding how to obtain both (nearly) optimal and efficient algorithms for the online eigenvector problem.

****

Near-Optimal Design of Experiments via Regret Minimization

Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, Yining Wang

We consider computationally tractable methods for the experimental design problem, where k out of n design points of dimension p are selected so that certain optimality criteria are approximately satisfied. Our algorithm finds a $(1+\epsilon)$-approximate optimal design when k is a linear function of p; in contrast, existing results require k to be super-linear in p. Our algorithm also handles all popular optimality criteria, while existing ones only handle one or two such criteria. Numerical results on synthetic and real-world design problems verify the practical effectiveness of the proposed algorithm.

****

OptNet: Differentiable Optimization as a Layer in Neural Networks

Brandon Amos, J. Zico Kolter

This paper presents OptNet, a network architecture that integrates optimization problems (here, specifically in the form of quadratic programs) as individual layers in larger end-to-end trainable deep networks. These layers encode constraints and complex dependencies between the hidden states that traditional convolutional and fully-connected layers often cannot capture. In this paper, we explore the foundations for such an architecture: we show how techniques from sensitivity analysis, bilevel optimization, and implicit differentiation can be used to exactly differentiate through these layers and with respect to layer parameters; we develop a highly efficient solver for these layers that exploits fast GPU-based batch solves within a primal-dual interior point method, and which provides backpropagation gradients with virtually no additional cost on top of the solve; and we highlight the application of these approaches in several problems. In one notable example, we show that the method is capable of learning to play mini-Sudoku (4x4) given just input and output games, with no a priori information about the rules of the game; this highlights the ability of our architecture to learn hard constraints better than other neural architectures.

****

Input Convex Neural Networks

Brandon Amos, Lei Xu, J. Zico Kolter

This paper presents the input convex neural network architecture. These are scalar-valued (potentially deep) neural networks with constraints on the network parameters such that the output of the network is a convex function of (some of) the inputs. The networks allow for efficient inference via optimization over some inputs to the network given others, and can be applied to settings including structured prediction, data imputation, reinforcement learning, and others. In this paper we lay the basic groundwork for these models, proposing methods for inference, optimization and learning, and analyze their representational power. We show that many existing neural network architectures can be made input-convex with a minor modification, and develop specialized optimization algorithms tailored to this setting. Finally, we highlight the performance of the methods on multi-label prediction, image completion, and reinforcement learning problems, where we show improvement over the existing state of the art in many cases.

****

An Efficient, Sparsity-Preserving, Online Algorithm for Low-Rank Approximation

David Anderson, Ming Gu

Low-rank matrix approximation is a fundamental tool in data analysis for processing large datasets, reducing noise, and finding important signals. In this work, we present a novel truncated LU factorization called Spectrum-Revealing LU (SRLU) for effective low-rank matrix approximation, and develop a fast algorithm to compute an SRLU factorization. We provide both matrix and singular value approximation error bounds for the SRLU approximation computed by our algorithm. Our analysis suggests that SRLU is competitive with the best low-rank matrix approximation methods, deterministic or randomized, in both computational complexity and approximation quality. Numeric experiments illustrate that SRLU preserves sparsity, highlights important data features and variables, can be efficiently updated, and calculates data approximations nearly as accurately as the best possible. To the best of our knowledge this is the first practical variant of the LU factorization for effective and efficient low-rank matrix approximation.

****

Modular Multitask Reinforcement Learning with Policy Sketches

Jacob Andreas, Dan Klein, Sergey Levine

We describe a framework for multitask deep reinforcement learning guided by policy sketches. Sketches annotate tasks with sequences of named subtasks, providing information about high-level structural relationships among tasks but not how to implement them—specifically not providing the detailed guidance used by much previous work on learning policy abstractions for RL (e.g. intermediate rewards, subtask completion signals, or intrinsic motivations). To learn from sketches, we present a model that associates every subtask with a modular subpolicy, and jointly maximizes reward over full task-specific policies by tying parameters across shared subpolicies. Optimization is accomplished via a decoupled actor–critic training objective that facilitates learning common behaviors from multiple dissimilar reward functions. We evaluate the effectiveness of our approach in three environments featuring both discrete and continuous control, and with sparse rewards that can be obtained only after completing a number of high-level subgoals. Experiments show that using our approach to learn policies guided by sketches gives better performance than existing techniques for learning task-specific or shared policies, while naturally inducing a library of interpretable primitive behaviors that can be recombined to rapidly adapt to new tasks.

****

Averaged-DQN: Variance Reduction and Stabilization for Deep Reinforcement Learning

Oron Anschel, Nir Baram, Nahum Shimkin

Instability and variability of Deep Reinforcement Learning (DRL) algorithms tend to adversely affect their performance. Averaged-DQN is a simple extension to the DQN algorithm, based on averaging previously learned Q-values estimates, which leads to a more stable training procedure and improved performance by reducing approximation error variance in the target values. To understand the effect of the algorithm, we examine the source of value function estimation errors and provide an analytical comparison within a simplified model. We further present experiments on the Arcade Learning Environment benchmark that demonstrate significantly improved stability and performance due to the proposed extension.

****

A Simple Multi-Class Boosting Framework with Theoretical Guarantees and Empirical Proficiency

Ron Appel, Pietro Perona

There is a need for simple yet accurate white-box learning systems that train quickly and with little data. To this end, we showcase REBEL, a multi-class boosting method, and present a novel family of weak learners called localized similarities. Our framework provably minimizes the training error of any dataset at an exponential rate. We carry out experiments on a variety of synthetic and real datasets, demonstrating a consistent tendency to avoid overfitting. We evaluate our method on MNIST and standard UCI datasets against other state-of-the-art method s, showing the empirical proficiency of our method.

****

Deep Voice: Real-time Neural Text-to-Speech

Sercan Ö. Ar■k, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, Mohammad Shoeybi

We present Deep Voice, a production-quality text-to-speech system constructed entirely from deep neural networks. Deep Voice lays the groundwork for truly end-to-end neural speech synthesis. The system comprises five major building blocks: a segmentation model for locating phoneme boundaries, a grapheme-to-phoneme conversion model, a phoneme duration prediction model, a fundamental frequency prediction model, and an audio synthesis model. For the segmentation model, we propose a novel way of performing phoneme boundary detection with deep neural networks using connectionist temporal classification (CTC) loss. For the audio synthesis model, we implement a variant of WaveNet that requires fewer parameters and trains faster than the original. By using a neural network for each component, our system is simpler and more flexible than traditional text-to-speech systems, where each component requires laborious feature engineering and extensive domain expertise. Finally, we show that inference with our system can be performed faster than real time and describe optimized WaveNet inference kernels on both CPU and GPU that achieve up to 400x speedups over existing implementations.

****

Oracle Complexity of Second-Order Methods for Finite-Sum Problems

Yossi Arjevani, Ohad Shamir

Finite-sum optimization problems are ubiquitous in machine learning, and are commonly solved using first-order methods which rely on gradient computations. Recently, there has been growing interest in second-order methods, which rely on both gradients and Hessians. In principle, second-order methods can require much fewer iterations than first-order methods, and hold the promise for more efficient algorithms. Although computing and manipulating Hessians is prohibitive for high-dimensional problems in general, the Hessians of individual functions in finite-sum problems can often be efficiently computed, e.g. because they possess a low-rank structure. Can second-order information indeed be used to solve such problems more efficiently? In this paper, we provide evidence that the answer – perhaps surprisingly – is negative, at least in terms of worst-case guarantees. However, we also discuss what additional assumptions and algorithmic approaches might potentially circumvent this negative result.

****

Wasserstein Generative Adversarial Networks

Martin Arjovsky, Soumith Chintala, Léon Bottou

We introduce a new algorithm named WGAN, an alternative to traditional GAN training. In this new model, we show that we can improve the stability of learning, get rid of problems like mode collapse, and provide meaningful learning curves useful for debugging and hyperparameter searches. Furthermore, we show that the corresponding optimization problem is sound, and provide extensive theoretical work highlighting the deep connections to different distances between distributions.

****

Generalization and Equilibrium in Generative Adversarial Nets (GANs)

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, Yi Zhang

It is shown that training of generative adversarial network (GAN) may not have good generalization properties; e.g., training may appear successful but the trained distribution may be far from target distribution in standard metrics. However, generalization does occur for a weaker metric called neural net distance. It is also shown that an approximate pure equilibrium exists in the discriminator/generator game for a natural training objective (Wasserstein) when generator capacity and training set sizes are moderate. This existence of equilibrium inspires MIX+GAN protocol, which can be combined with any existing GAN training, and empirically shown to improve some of them.

****

A Closer Look at Memorization in Deep Networks
Devansh Arpit, Stanis■aw Jastrz■bski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, Simon Lacoste-Julien
We examine the role of memorization in deep learning, drawing connections to capacity, generalization, and adversarial robustness. While deep networks are capable of memorizing noise data, our results suggest that they tend to prioritize learning simple patterns first. In our experiments, we expose qualitative differences in gradient-based optimization of deep neural networks (DNNs) on noise vs.~real data. We also demonstrate that for appropriately tuned explicit regularization (e.g.,~dropout) we can degrade DNN training performance on noise datasets without compromising generalization on real data. Our analysis suggests that the notions of effective capacity which are dataset independent are unlikely to explain the generalization performance of deep networks when trained with gradient based methods because training data itself plays an important role in determining the degree of memorization.
****

An Alternative Softmax Operator for Reinforcement Learning
Kavosh Asadi, Michael L. Littman
A softmax operator applied to a set of values acts somewhat like the maximization function and somewhat like an average. In sequential decision making, softmax is often used in settings where it is necessary to maximize utility but also to hedge against problems that arise from putting all of one’s weight behind a single maximum utility decision. The Boltzmann softmax operator is the most commonly used softmax operator in this setting, but we show that this operator is prone to misbehavior. In this work, we study a differentiable softmax operator that, among other properties, is a non-expansion ensuring a convergent behavior in learning and planning. We introduce a variant of SARSA algorithm that, by utilizing the new operator, computes a Boltzmann policy with a state-dependent temperature parameter. We show that the algorithm is convergent and that it performs favorably in practice.
****

Random Fourier Features for Kernel Ridge Regression: Approximation Bounds and Statistical Guarantees
Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, Amir Zandieh
Random Fourier features is one of the most popular techniques for scaling up kernel methods, such as kernel ridge regression. However, despite impressive empirical results, the statistical properties of random Fourier features are still not well understood. In this paper we take steps toward filling this gap. Specifically, we approach random Fourier features from a spectral matrix approximation point of view, give tight bounds on the number of Fourier features required to achieve a spectral approximation, and show how spectral matrix approximation bounds imply statistical guarantees for kernel ridge regression.
****

Minimax Regret Bounds for Reinforcement Learning
Mohammad Gheshlaghi Azar, Ian Osband, Rémi Munos
We consider the problem of provably optimal exploration in reinforcement learning for finite horizon MDPs. We show that an optimistic modification to value iteration achieves a regret bound of $\tilde {O}(\sqrt{HSAT} + H^2S^2A+H\sqrt{T})$ where H is the time horizon, S the number of states, A the number of actions and T the number of time-steps. This result improves over the best previous known bound $\tilde {O}(HS \sqrt{AT})$ achieved by the UCRL2 algorithm. The key significance of our new results is that when $T\geq H^3S^3A$ and $SA\geq H$, it leads to a regret of $\tilde{O}(\sqrt{HSAT})$ that matches the established lower bound of $\Omega(\sqrt{HSAT})$ up to a logarithmic factor. Our analysis contain two key insights. We use careful application of concentration inequalities to the optimal value function as a whole, rather than to the transitions probabilities (to improve scaling in S), and we define Bernstein-based “exploration bonuses” that use the empirical variance of the estimated values at the next states

(to improve scaling in H).

****

Learning the Structure of Generative Models without Labeled Data

Stephen H. Bach, Bryan He, Alexander Ratner, Christopher Ré

Curating labeled training data has become the primary bottleneck in machine learning. Recent frameworks address this bottleneck with generative models to synthesize labels at scale from weak supervision sources. The generative model’s dependency structure directly affects the quality of the estimated labels, but selecting a structure automatically without any labeled data is a distinct challenge. We propose a structure estimation method that maximizes the l1-regularized marginal pseudolikelihood of the observed data. Our analysis shows that the amount of unlabeled data required to identify the true structure scales sublinearly in the number of possible dependencies for a broad class of models. Simulations show that our method is 100x faster than a maximum likelihood approach and selects 1/4 as many extraneous dependencies. We also show that our method provides an average of 1.5 F1 points of improvement over existing, user-developed information extraction applications on real-world data such as PubMed journal abstracts.

****

Uniform Deviation Bounds for k-Means Clustering

Olivier Bachem, Mario Lucic, S. Hamed Hassani, Andreas Krause

Uniform deviation bounds limit the difference between a model’s expected loss and its loss on an empirical sample uniformly for all models in a learning problem. In this paper, we provide a novel framework to obtain uniform deviation bounds for loss functions which are unbounded. As a result, we obtain competitive uniform deviation bounds for k-Means clustering under weak assumptions on the underlying distribution. If the fourth moment is bounded, we prove a rate of $O(m^{-1/2})$ compared to the previously known $O(m^{-1/4})$ rate. Furthermore, we show that the rate also depends on the kurtosis – the normalized fourth moment which measures the “tailedness” of a distribution. We also provide improved rates under progressively stronger assumptions, namely, bounded higher moments, subgaussianity and bounded support of the underlying distribution.

****

Distributed and Provably Good Seedings for k-Means in Constant Rounds

Olivier Bachem, Mario Lucic, Andreas Krause

The k-Means++ algorithm is the state of the art algorithm to solve k-Means clustering problems as the computed clusterings are O(log k) competitive in expectation. However, its seeding step requires k inherently sequential passes through the full data set making it hard to scale to massive data sets. The standard remedy is to use the k-Means|| algorithm which reduces the number of sequential rounds and is thus suitable for a distributed setting. In this paper, we provide a novel analysis of the k-Means|| algorithm that bounds the expected solution quality for any number of rounds and oversampling factors greater than k, the two parameters one needs to choose in practice. In particular, we show that k-Means|| provides provably good clusterings even for a small, constant number of iterations. This theoretical finding explains the common observation that k-Means|| performs extremely well in practice even if the number of rounds is low. We further provide a hard instance that shows that an additive error term as encountered in our analysis is inevitable if less than k-1 rounds are employed.

****

Learning Algorithms for Active Learning

Philip Bachman, Alessandro Sordoni, Adam Trischler

We introduce a model that learns active learning algorithms via metalearning. For a distribution of related tasks, our model jointly learns: a data representation, an item selection heuristic, and a prediction function. Our model uses the item selection heuristic to construct a labeled support set for training the prediction function. Using the Omniglot and MovieLens datasets, we test our model in synthetic and practical settings.

****

Improving Viterbi is Hard: Better Runtimes Imply Faster Clique Algorithms

Arturs Backurs, Christos Tzamos

The classic algorithm of Viterbi computes the most likely path in a Hidden Markov Model (HMM) that results in a given sequence of observations. It runs in time $O(Tn^2)$ given a sequence of T observations from a HMM with n states. Despite significant interest in the problem and prolonged effort by different communities, no known algorithm achieves more than a polylogarithmic speedup. In this paper, we explain this difficulty by providing matching conditional lower bounds. Our lower bounds are based on assumptions that the best known algorithms for the All-Pairs Shortest Paths problem (APSP) and for the Max-Weight k-Clique problem in edge-weighted graphs are essentially tight. Finally, using a recent algorithm by Green Larsen and Williams for online Boolean matrix-vector multiplication, we get a $2^{\Omega(\sqrt{\log n})}$ speedup for the Viterbi algorithm when there are few distinct transition probabilities in the HMM.

****

Differentially Private Clustering in High-Dimensional Euclidean Spaces

Maria-Florina Balcan, Travis Dick, Yingyu Liang, Wenlong Mou, Hongyang Zhang

We study the problem of clustering sensitive data while preserving the privacy of individuals represented in the dataset, which has broad applications in practical machine learning and data analysis tasks. Although the problem has been widely studied in the context of low-dimensional, discrete spaces, much remains unknown concerning private clustering in high-dimensional Euclidean spaces $\mathbb{R}^d$. In this work, we give differentially private and efficient algorithms achieving strong guarantees for k-means and k-median clustering when $d=\Omega(\mathsf{polylog}(n))$. Our algorithm achieves clustering loss at most $\log^3(n)\mathsf{OPT}+\mathsf{poly}(\log n,d,k)$, advancing the state-of-the-art result of $\sqrt{d}\mathsf{OPT}+\mathsf{poly}(\log n,d^d,k^d)$. We also study the case where the data points are s-sparse and show that the clustering loss can scale logarithmically with d, i.e., $\log^3(n)\mathsf{OPT}+\mathsf{poly}(\log n,\log d,k,s)$. Experiments on both synthetic and real datasets verify the effectiveness of the proposed method.

****

Strongly-Typed Agents are Guaranteed to Interact Safely

David Balduzzi

As artificial agents proliferate, it is becoming increasingly important to ensure that their interactions with one another are well-behaved. In this paper, we formalize a common-sense notion of when algorithms are well-behaved: an algorithm is safe if it does no harm. Motivated by recent progress in deep learning, we focus on the specific case where agents update their actions according to gradient descent. The paper shows that that gradient descent converges to a Nash equilibrium in safe games. The main contribution is to define strongly-typed agents and show they are guaranteed to interact safely, thereby providing sufficient conditions to guarantee safe interactions. A series of examples show that strong-typing generalizes certain key features of convexity, is closely related to blind source separation, and introduces a new perspective on classical multilinear games based on tensor decomposition.

****

The Shattered Gradients Problem: If resnets are the answer, then what is the question?

David Balduzzi, Marcus Frean, Lennox Leary, J. P. Lewis, Kurt Wan-Duo Ma, Brian McWilliams

A long-standing obstacle to progress in deep learning is the problem of vanishing and exploding gradients. Although, the problem has largely been overcome via carefully constructed initializations and batch normalization, architectures incorporating skip-connections such as highway and resnets perform much better than standard feedforward architectures despite well-chosen initialization and batch normalization. In this paper, we identify the shattered gradients problem. Specifically, we show that the correlation between gradients in standard feedforward networks decays exponentially with depth resulting in gradients that resemble white noise whereas, in contrast, the gradients in architectures with skip-connections are far more resistant to shattering, decaying sublinearly. Detailed empirical evidence is presented in support of the analysis, on both fully-connected ne tworks and convnets. Finally, we present a new “looks linear” (LL) initialization that prevents shattering, with preliminary experiments showing the new initialization allows to train very deep networks without the addition of skip-connections.

****

Neural Taylor Approximations: Convergence and Exploration in Rectifier Networks

David Balduzzi, Brian McWilliams, Tony Butler-Yeoman

Modern convolutional networks, incorporating rectifiers and max-pooling, are neither smooth nor convex; standard guarantees therefore do not apply. Nevertheless, methods from convex optimization such as gradient descent and Adam are widely used as building blocks for deep learning algorithms. This paper provides the first convergence guarantee applicable to modern convnets, which furthermore matches a lower bound for convex nonsmooth functions. The key technical tool is the neural Taylor approximation – a straightforward application of Taylor expansions to neural networks – and the associated Taylor loss. Experiments on a range of optimizers, layers, and tasks provide evidence that the analysis accurately captures the dynamics of neural optimization. The second half of the paper applies the Taylor approximation to isolate the main difficulty in training rectifier nets – that gradients are shattered – and investigates the hypothesis that, by exploring the space of activation configurations more thoroughly, adaptive optimizers such as RMSProp and Adam are able to converge to better solutions.

****

Spectral Learning from a Single Trajectory under Finite-State Policies

Borja Balle, Odalric-Ambrym Maillard

We present spectral methods of moments for learning sequential models from a single trajectory, in stark contrast with the classical literature that assumes the availability of multiple i.i.d. trajectories. Our approach leverages an efficient SVD-based learning algorithm for weighted automata and provides the first rigorous analysis for learning many important models using dependent data. We state and analyze the algorithm under three increasingly difficult scenarios: probabilistic automata, stochastic weighted automata, and reactive predictive state representations controlled by a finite-state policy. Our proofs include novel tools for studying mixing properties of stochastic weighted automata.

****

Lost Relatives of the Gumbel Trick

Matej Balog, Nilesh Tripuraneni, Zoubin Ghahramani, Adrian Weller

The Gumbel trick is a method to sample from a discrete probability distribution, or to estimate its normalizing partition function. The method relies on repeatedly applying a random perturbation to the distribution in a particular way, each time solving for the most likely configuration. We derive an entire family of related methods, of which the Gumbel trick is one member, and show that the new methods have superior properties in several settings with minimal additional computational cost. In particular, for the Gumbel trick to yield computational benefits for discrete graphical models, Gumbel perturbations on all configurations are typically replaced with so-called low-rank perturbations. We show how a subfamily of our new methods adapts to this setting, proving new upper and lower bounds on the log partition function and deriving a family of sequential samplers for the Gibbs distribution. Finally, we balance the discussion by showing how the simpler analytical form of the Gumbel trick enables additional theoretical results.

****

Dynamic Word Embeddings

Robert Bamler, Stephan Mandt

We present a probabilistic language model for time-stamped text data which tracks the semantic evolution of individual words over time. The model represents words and contexts by latent trajectories in an embedding space. At each moment in time, the embedding vectors are inferred from a probabilistic version of word2vec [Mikolov et al., 2013]. These embedding vectors are connected in time through a latent diffusion process. We describe two scalable variational inference algorithms–skip-gram smoothing and skip-gram filtering–that allow us to train the mod

el jointly over all times; thus learning on all data while simultaneously allowing word and context vectors to drift. Experimental results on three different corpora demonstrate that our dynamic model infers word embedding trajectories that are more interpretable and lead to higher predictive likelihoods than competing methods that are based on static models trained separately on time slices.
****

End-to-End Differentiable Adversarial Imitation Learning
Nir Baram, Oron Anschel, Itai Caspi, Shie Mannor

Generative Adversarial Networks (GANs) have been successfully applied to the problem of policy imitation in a model-free setup. However, the computation graph of GANs, that include a stochastic policy as the generative model, is no longer differentiable end-to-end, which requires the use of high-variance gradient estimation. In this paper, we introduce the Model-based Generative Adversarial Imitation Learning (MGAIL) algorithm. We show how to use a forward model to make the computation fully differentiable, which enables training policies using the exact gradient of the discriminator. The resulting algorithm trains competent policies using relatively fewer expert samples and interactions with the environment. We test it on both discrete and continuous action domains and report results that surpass the state-of-the-art.
****

Emulating the Expert: Inverse Optimization through Online Learning
Andreas Bärmann, Sebastian Pokutta, Oskar Schneider

In this paper, we demonstrate how to learn the objective function of a decision maker while only observing the problem input data and the decision maker’s corresponding decisions over multiple rounds. Our approach is based on online learning techniques and works for linear objectives over arbitrary sets for which we have a linear optimization oracle and as such generalizes previous work based on KKT-system decomposition and dualization approaches. The applicability of our framework for learning linear constraints is also discussed briefly. Our algorithm converges at a rate of O(1/sqrt(T)), and we demonstrate its effectiveness and applications in preliminary computational results.
****

Unimodal Probability Distributions for Deep Ordinal Classification
Christopher Beckham, Christopher Pal

Probability distributions produced by the cross-entropy loss for ordinal classification problems can possess undesired properties. We propose a straightforward technique to constrain discrete ordinal probability distributions to be unimodal via the use of the Poisson and binomial probability distributions. We evaluate this approach in the context of deep learning on two large ordinal image datasets, obtaining promising results.
****

Globally Induced Forest: A Prepruning Compression Scheme
Jean-Michel Begon, Arnaud Joly, Pierre Geurts

Tree-based ensemble models are heavy memory-wise. An undesired state of affairs considering nowadays datasets, memory-constrained environment and fitting/prediction times. In this paper, we propose the Globally Induced Forest (GIF) to remedy this problem. GIF is a fast prepruning approach to build lightweight ensembles by iteratively deepening the current forest. It mixes local and global optimizations to produce accurate predictions under memory constraints in reasonable time. We show that the proposed method is more than competitive with standard tree-based ensembles under corresponding constraints, and can sometimes even surpass much larger models.
****

End-to-End Learning for Structured Prediction Energy Networks
David Belanger, Bishan Yang, Andrew McCallum

Structured Prediction Energy Networks (SPENs) are a simple, yet expressive family of structured prediction models (Belanger and McCallum, 2016). An energy function over candidate structured outputs is given by a deep network, and predictions are formed by gradient-based optimization. This paper presents end-to-end learning for SPENs, where the energy function is discriminatively trained by back-pr opagating through gradient-based prediction. In our experience, the approach is substantially more accurate than the structured SVM method of Belanger and McCallum (2016), as it allows us to use more sophisticated non-convex energies. We provide a collection of techniques for improving the speed, accuracy, and memory requirements of end-to-end SPENs, and demonstrate the power of our method on 7-Scenes image denoising and CoNLL-2005 semantic role labeling tasks. In both, inexact minimization of non-convex SPEN energies is superior to baseline methods that use simplistic energy functions that can be minimized exactly.

****

Learning to Discover Sparse Graphical Models

Eugene Belilovsky, Kyle Kastner, Gael Varoquaux, Matthew B. Blaschko

We consider structure discovery of undirected graphical models from observational data. Inferring likely structures from few examples is a complex task often requiring the formulation of priors and sophisticated inference procedures. Popular methods rely on estimating a penalized maximum likelihood of the precision matrix. However, in these approaches structure recovery is an indirect consequence of the data-fit term, the penalty can be difficult to adapt for domain-specific knowledge, and the inference is computationally demanding. By contrast, it may be easier to generate training samples of data that arise from graphs with the desired structure properties. We propose here to leverage this latter source of information as training data to learn a function, parametrized by a neural network, that maps empirical covariance matrices to estimated graph structures. Learning this function brings two benefits: it implicitly models the desired structure or sparsity properties to form suitable priors, and it can be tailored to the specific problem of edge structure discovery, rather than maximizing data likelihood. Applying this framework, we find our learnable graph-discovery method trained on synthetic data generalizes well: identifying relevant edges in both synthetic and real data, completely unknown at training time. We find that on genetics, brain imaging, and simulation data we obtain performance generally superior to analytical methods.

****

A Distributional Perspective on Reinforcement Learning

Marc G. Bellemare, Will Dabney, Rémi Munos

In this paper we argue for the fundamental importance of the value distribution: the distribution of the random return received by a reinforcement learning agent. This is in contrast to the common approach to reinforcement learning which models the expectation of this return, or value. Although there is an established body of literature studying the value distribution, thus far it has always been used for a specific purpose such as implementing risk-aware behaviour. We begin with theoretical results in both the policy evaluation and control settings, exposing a significant distributional instability in the latter. We then use the distributional perspective to design a new algorithm which applies Bellman's equation to the learning of approximate value distributions. We evaluate our algorithm using the suite of games from the Arcade Learning Environment. We obtain both state-of-the-art results and anecdotal evidence demonstrating the importance of the value distribution in approximate reinforcement learning. Finally, we combine theoretical and empirical evidence to highlight the ways in which the value distribution impacts learning in the approximate setting.

****

Neural Optimizer Search with Reinforcement Learning

Irwan Bello, Barret Zoph, Vijay Vasudevan, Quoc V. Le

We present an approach to automate the process of discovering optimization methods, with a focus on deep learning architectures. We train a Recurrent Neural Network controller to generate a string in a specific domain language that describes a mathematical update equation based on a list of primitive functions, such as the gradient, running average of the gradient, etc. The controller is trained with Reinforcement Learning to maximize the performance of a model after a few epochs. On CIFAR-10, our method discovers several update rules that are better than many commonly used optimizers, such as Adam, RMSProp, or SGD with and without Momentum on a ConvNet model. These optimizers can also be transferred to perform well on different neural network architectures, including Google’s neural machine translation system.

****

Learning Texture Manifolds with the Periodic Spatial GAN

Urs Bergmann, Nikolay Jetchev, Roland Vollgraf

This paper introduces a novel approach to texture synthesis based on generative adversarial networks (GAN) (Goodfellow et al., 2014), and call this technique Periodic Spatial GAN (PSGAN). The PSGAN has several novel abilities which surpass the current state of the art in texture synthesis. First, we can learn multiple textures, periodic or non-periodic, from datasets of one or more complex large images. Second, we show that the image generation with PSGANs has properties of a texture manifold: we can smoothly interpolate between samples in the structured noise space and generate novel samples, which lie perceptually between the textures of the original dataset. We make multiple experiments which show that PSGANs can flexibly handle diverse texture and image data sources, and the method is highly scalable and can generate output images of arbitrary large size.

****

Differentially Private Learning of Undirected Graphical Models Using Collective Graphical Models

Garrett Bernstein, Ryan McKenna, Tao Sun, Daniel Sheldon, Michael Hay, Gerome Miklau

We investigate the problem of learning discrete graphical models in a differentially private way. Approaches to this problem range from privileged algorithms that conduct learning completely behind the privacy barrier to schemes that release private summary statistics paired with algorithms to learn parameters from those statistics. We show that the approach of releasing noisy sufficient statistics using the Laplace mechanism achieves a good trade-off between privacy, utility, and practicality. A naive learning algorithm that uses the noisy sufficient statistics “as is” outperforms general-purpose differentially private learning algorithms. However, it has three limitations: it ignores knowledge about the data generating process, rests on uncertain theoretical foundations, and exhibits certain pathologies. We develop a more principled approach that applies the formalism of collective graphical models to perform inference over the true sufficient statistics within an expectation-maximization framework. We show that this learns better models than competing approaches on both synthetic data and on real human mobility data used as a case study.

****

Efficient Online Bandit Multiclass Learning with $\tilde{O}(\sqrt{T})$ Regret

Alina Beygelzimer, Francesco Orabona, Chicheng Zhang

We present an efficient second-order algorithm with $\tilde{O}(1/\eta \sqrt{T})$ regret for the bandit online multiclass problem. The regret bound holds simultaneously with respect to a family of loss functions parameterized by η, ranging from hinge loss ($\eta=0$) to squared hinge loss ($\eta=1$). This provides a solution to the open problem of (Abernethy, J. and Rakhlin, A. An efficient bandit algorithm for $\sqrt{T}$-regret in online multiclass prediction? In COLT, 2009). We test our algorithm experimentally, showing that it performs favorably against earlier algorithms.

****

Guarantees for Greedy Maximization of Non-submodular Functions with Applications

Andrew An Bian, Joachim M. Buhmann, Andreas Krause, Sebastian Tschiatschek

We investigate the performance of the standard Greedy algorithm for cardinality constrained maximization of non-submodular nondecreasing set functions. While there are strong theoretical guarantees on the performance of Greedy for maximizing submodular functions, there are few guarantees for non-submodular ones. However, Greedy enjoys strong empirical performance for many important non-submodular functions, e.g., the Bayesian A-optimality objective in experimental design. We prove theoretical guarantees supporting the empirical performance. Our guarantees are characterized by a combination of the (generalized) curvature α and the submodularity ratio γ. In particular, we prove that Greedy enjoys a tight approximation guarantee of $\frac{1}{\alpha}(1- e^{-\gamma\alpha})$ for ca

rdinality constrained maximization. In addition, we bound the submodularity ratio and curvature for several important real-world objectives, including the Bayesian A-optimality objective, the determinantal function of a square submatrix and certain linear programs with combinatorial constraints. We experimentally validate our theoretical findings for both synthetic and real-world applications.
****

Robust Submodular Maximization: A Non-Uniform Partitioning Approach
Ilija Bogunovic, Slobodan Mitrovi■, Jonathan Scarlett, Volkan Cevher
We study the problem of maximizing a monotone submodular function subject to a cardinality constraint k, with the added twist that a number of items τ from the returned set may be removed. We focus on the worst-case setting considered by Orlin et al.\ (2016), in which a constant-factor approximation guarantee was given for $\tau = o(\sqrt{k})$. In this paper, we solve a key open problem raised therein, presenting a new Partitioned Robust (PRo) submodular maximization algorithm that achieves the same guarantee for more general $\tau = o(k)$. Our algorithm constructs partitions consisting of buckets with exponentially increasing sizes, and applies standard submodular optimization subroutines on the buckets in order to construct the robust solution. We numerically demonstrate the performance of PRo in data summarization and influence maximization, demonstrating gains over both the greedy algorithm and the algorithm of Orlin et al.\ (2016).
****

Unsupervised Learning by Predicting Noise
Piotr Bojanowski, Armand Joulin
Convolutional neural networks provide visual features that perform remarkably well in many computer vision applications. However, training these networks requires significant amounts of supervision; this paper introduces a generic framework to train such networks, end-to-end, with no supervision. We propose to fix a set of target representations, called Noise As Targets (NAT), and to constrain the deep features to align to them. This domain agnostic approach avoids the standard unsupervised learning issues of trivial solutions and collapsing of the features. Thanks to a stochastic batch reassignment strategy and a separable square loss function, it scales to millions of images. The proposed approach produces representations that perform on par with the state-of-the-arts among unsupervised methods on ImageNet and Pascal VOC.
****

Adaptive Neural Networks for Efficient Inference
Tolga Bolukbasi, Joseph Wang, Ofer Dekel, Venkatesh Saligrama
We present an approach to adaptively utilize deep neural networks in order to reduce the evaluation time on new examples without loss of accuracy. Rather than attempting to redesign or approximate existing networks, we propose two schemes that adaptively utilize networks. We first pose an adaptive network evaluation scheme, where we learn a system to adaptively choose the components of a deep network to be evaluated for each example. By allowing examples correctly classified using early layers of the system to exit, we avoid the computational time associated with full evaluation of the network. We extend this to learn a network selection system that adaptively selects the network to be evaluated for each example. We show that computational time can be dramatically reduced by exploiting the fact that many examples can be correctly classified using relatively efficient networks and that complex, computationally costly networks are only necessary for a small fraction of examples. We pose a global objective for learning an adaptive early exit or network selection policy and solve it by reducing the policy learning problem to a layer-by-layer weighted binary classification problem. Empirically, these approaches yield dramatic reductions in computational cost, with up to a 2.8x speedup on state-of-the-art networks from the ImageNet image recognition challenge with minimal ($<1\%$) loss of top5 accuracy.
****

Compressed Sensing using Generative Models
Ashish Bora, Ajil Jalal, Eric Price, Alexandros G. Dimakis
The goal of compressed sensing is to estimate a vector from an underdetermined system of noisy linear measurements, by making use of prior knowledge on the stru

cture of vectors in the relevant domain. For almost all results in this literature, the structure is represented by sparsity in a well-chosen basis. We show how to achieve guarantees similar to standard compressed sensing but without employing sparsity at all. Instead, we suppose that vectors lie near the range of a generative model $G: \mathbb{R}^k \to \mathbb{R}^n$. Our main theorem is that, if G is L-Lipschitz, then roughly $\mathcal{O}(k \log L)$ random Gaussian measurements suffice for an ℓ_2/ℓ_2 recovery guarantee. We demonstrate our results using generative models from published variational autoencoder and generative adversarial networks. Our method can use 5-10x fewer measurements than Lasso for the same accuracy.

****

Programming with a Differentiable Forth Interpreter

Matko Bošnjak, Tim Rocktäschel, Jason Naradowsky, Sebastian Riedel

Given that in practice training data is scarce for all but a small set of problems, a core question is how to incorporate prior knowledge into a model. In this paper, we consider the case of prior procedural knowledge for neural networks, such as knowing how a program should traverse a sequence, but not what local actions should be performed at each step. To this end, we present an end-to-end differentiable interpreter for the programming language Forth which enables programmers to write program sketches with slots that can be filled with behaviour trained from program input-output data. We can optimise this behaviour directly through gradient descent techniques on user-specified objectives, and also integrate the program into any larger neural computation graph. We show empirically that our interpreter is able to effectively leverage different levels of prior program structure and learn complex behaviours such as sequence sorting and addition. When connected to outputs of an LSTM and trained jointly, our interpreter achieves state-of-the-art accuracy for end-to-end reasoning about quantities expressed in natural language stories.

****

Practical Gauss-Newton Optimisation for Deep Learning

Aleksandar Botev, Hippolyt Ritter, David Barber

We present an efficient block-diagonal approximation to the Gauss-Newton matrix for feedforward neural networks. Our resulting algorithm is competitive against state-of-the-art first-order optimisation methods, with sometimes significant improvement in optimisation performance. Unlike first-order methods, for which hyperparameter tuning of the optimisation parameters is often a laborious process, our approach can provide good performance even when used with default settings. A side result of our work is that for piecewise linear transfer functions, the network objective function can have no differentiable local maxima, which may partially explain why such transfer functions facilitate effective optimisation.

****

Lazifying Conditional Gradient Algorithms

Gábor Braun, Sebastian Pokutta, Daniel Zink

Conditional gradient algorithms (also often called Frank-Wolfe algorithms) are popular due to their simplicity of only requiring a linear optimization oracle and more recently they also gained significant traction for online learning. While simple in principle, in many cases the actual implementation of the linear optimization oracle is costly. We show a general method to lazify various conditional gradient algorithms, which in actual computations leads to several orders of magnitude of speedup in wall-clock time. This is achieved by using a faster separation oracle instead of a linear optimization oracle, relying only on few linear optimization oracle calls.

****

Clustering High Dimensional Dynamic Data Streams

Vladimir Braverman, Gereon Frahling, Harry Lang, Christian Sohler, Lin F. Yang

We present data streaming algorithms for the k-median problem in high-dimensional dynamic geometric data streams, i.e. streams allowing both insertions and deletions of points from a discrete Euclidean space $\{1, 2, \ldots \Delta\}^d$. Our algorithms use $k \epsilon^{-2} \mathrm{poly}(d \log \Delta)$ space/time and maintain with high probability a small weighted set of points (a coreset) such t hat for every set of k centers the cost of the coreset $(1+\epsilon)$-approximates the cost of the streamed point set. We also provide algorithms that guarantee only positive weights in the coreset with additional logarithmic factors in the space and time complexities. We can use this positively-weighted coreset to compute a $(1+\epsilon)$-approximation for the k-median problem by any efficient offline k-median algorithm. All previous algorithms for computing a $(1+\epsilon)$-approximation for the k-median problem over dynamic data streams required space and time exponential in d. Our algorithms can be generalized to metric spaces of bounded doubling dimension.

****

On the Sampling Problem for Kernel Quadrature

François-Xavier Briol, Chris J. Oates, Jon Cockayne, Wilson Ye Chen, Mark Girolami

The standard Kernel Quadrature method for numerical integration with random point sets (also called Bayesian Monte Carlo) is known to converge in root mean square error at a rate determined by the ratio s/d, where s and d encode the smoothness and dimension of the integrand. However, an empirical investigation reveals that the rate constant C is highly sensitive to the distribution of the random points. In contrast to standard Monte Carlo integration, for which optimal importance sampling is well-understood, the sampling distribution that minimises C for Kernel Quadrature does not admit a closed form. This paper argues that the practical choice of sampling distribution is an important open problem. One solution is considered; a novel automatic approach based on adaptive tempering and sequential Monte Carlo. Empirical results demonstrate a dramatic reduction in integration error of up to 4 orders of magnitude can be achieved with the proposed method.

****

Reduced Space and Faster Convergence in Imperfect-Information Games via Pruning

Noam Brown, Tuomas Sandholm

Iterative algorithms such as Counterfactual Regret Minimization (CFR) are the most popular way to solve large zero-sum imperfect-information games. In this paper we introduce Best-Response Pruning (BRP), an improvement to iterative algorithms such as CFR that allows poorly-performing actions to be temporarily pruned. We prove that when using CFR in zero-sum games, adding BRP will asymptotically prune any action that is not part of a best response to some Nash equilibrium. This leads to provably faster convergence and lower space requirements. Experiments show that BRP results in a factor of 7 reduction in space, and the reduction factor increases with game size.

****

Globally Optimal Gradient Descent for a ConvNet with Gaussian Inputs

Alon Brutzkus, Amir Globerson

Deep learning models are often successfully trained using gradient descent, despite the worst case hardness of the underlying non-convex optimization problem. The key question is then under what conditions can one prove that optimization will succeed. Here we provide a strong result of this kind. We consider a neural net with one hidden layer and a convolutional structure with no overlap and a ReLU activation function. For this architecture we show that learning is NP-complete in the general case, but that when the input distribution is Gaussian, gradient descent converges to the global optimum in polynomial time. To the best of our knowledge, this is the first global optimality guarantee of gradient descent on a convolutional neural network with ReLU activations.

****

Deep Tensor Convolution on Multicores

David Budden, Alexander Matveev, Shibani Santurkar, Shraman Ray Chaudhuri, Nir Shavit

Deep convolutional neural networks (ConvNets) of 3-dimensional kernels allow joint modeling of spatiotemporal features. These networks have improved performance of video and volumetric image analysis, but have been limited in size due to the low memory ceiling of GPU hardware. Existing CPU implementations overcome this constraint but are impractically slow. Here we extend and optimize the faster W inograd-class of convolutional algorithms to the N-dimensional case and specifically for CPU hardware. First, we remove the need to manually hand-craft algorithms by exploiting the relaxed constraints and cheap sparse access of CPU memory. Second, we maximize CPU utilization and multicore scalability by transforming data matrices to be cache-aware, integer multiples of AVX vector widths. Treating 2-dimensional ConvNets as a special (and the least beneficial) case of our approach, we demonstrate a 5 to 25-fold improvement in throughput compared to previous state-of-the-art.

****

Multi-objective Bandits: Optimizing the Generalized Gini Index

Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, Shie Mannor

We study the multi-armed bandit (MAB) problem where the agent receives a vectorial feedback that encodes many possibly competing objectives to be optimized. The goal of the agent is to find a policy, which can optimize these objectives simultaneously in a fair way. This multi-objective online optimization problem is formalized by using the Generalized Gini Index (GGI) aggregation function. We propose an online gradient descent algorithm which exploits the convexity of the GGI aggregation function, and controls the exploration in a careful way achieving a distribution-free regret $\tilde{O}(T^{-1/2})$ with high probability. We test our algorithm on synthetic data as well as on an electric battery control problem where the goal is to trade off the use of the different cells of a battery in order to balance their respective degradation rates.

****

Priv’IT: Private and Sample Efficient Identity Testing

Bryan Cai, Constantinos Daskalakis, Gautam Kamath

We develop differentially private hypothesis testing methods for the small sample regime. Given a sample $\mathcal{D}$ from a categorical distribution p over some domain Σ, an explicitly described distribution q over Σ, some privacy parameter ϵ, accuracy parameter α, and requirements β_I and β_II for the type I and type II errors of our test, the goal is to distinguish between $p=q$ and $d_\mathrm{tv}(p,q) \ge \alpha$. We provide theoretical bounds for the sample size $|\mathcal{D}|$ so that our method both satisfies $(\epsilon,0)$-differential privacy, and guarantees β_I and β_II type I and type II errors. We show that differential privacy may come for free in some regimes of parameters, and we always beat the sample complexity resulting from running the χ^2-test with noisy counts, or standard approaches such as repetition for endowing non-private χ^2-style statistics with differential privacy guarantees. We experimentally compare the sample complexity of our method to that of recently proposed methods for private hypothesis testing.

****

Second-Order Kernel Online Convex Optimization with Adaptive Sketching

Daniele Calandriello, Alessandro Lazaric, Michal Valko

Kernel online convex optimization (KOCO) is a framework combining the expressiveness of non-parametric kernel models with the regret guarantees of online learning. First-order KOCO methods such as functional gradient descent require only $O(t)$ time and space per iteration, and, when the only information on the losses is their convexity, achieve a minimax optimal $O(\sqrt{T})$ regret. Nonetheless, many common losses in kernel problems, such as squared loss, logistic loss, and squared hinge loss posses stronger curvature that can be exploited. In this case, second-order KOCO methods achieve $O(\log(\mathrm{Det}(K)))$ regret, which we show scales as $O(deff \log T)$, where $deff$ is the effective dimension of the problem and is usually much smaller than $O(\sqrt{T})$. The main drawback of second-order methods is their much higher $O(t^2)$ space and time complexity. In this paper, we introduce kernel online Newton step (KONS), a new second-order KOCO method that also achieves $O(deff\log T)$ regret. To address the computational complexity of second-order methods, we introduce a new matrix sketching algorithm for the kernel matrix~K, and show that for a chosen parameter $\gamma \leq 1$ our Sketched-KONS reduces the space and time complexity by a factor of γ^2 to $O(t^2\gamma^2)$ space and time per iteration, while incurring only $1/\

gamma$ times more regret.

****

“Convex Until Proven Guilty”: Dimension-Free Acceleration of Gradient Descent on Non-Convex Functions

Yair Carmon, John C. Duchi, Oliver Hinder, Aaron Sidford

We develop and analyze a variant of Nesterov’s accelerated gradient descent (AGD) for minimization of smooth non-convex functions. We prove that one of two cases occurs: either our AGD variant converges quickly, as if the function was convex, or we produce a certificate that the function is “guilty” of being non-convex. This non-convexity certificate allows us to exploit negative curvature and obtain deterministic, dimension-free acceleration of convergence for non-convex functions. For a function f with Lipschitz continuous gradient and Hessian, we compute a point x with $\|\nabla f(x)\| \le \epsilon$ in $O(\epsilon^{-7/4} \log(1/ \epsilon))$ gradient and function evaluations. Assuming additionally that the third derivative is Lipschitz, we require only $O(\epsilon^{-5/3} \log(1/ \epsilon))$ evaluations.

****

Sliced Wasserstein Kernel for Persistence Diagrams

Mathieu Carrière, Marco Cuturi, Steve Oudot

Persistence diagrams (PDs) play a key role in topological data analysis (TDA), in which they are routinely used to describe succinctly complex topological properties of complicated shapes. PDs enjoy strong stability properties and have proven their utility in various learning contexts. They do not, however, live in a space naturally endowed with a Hilbert structure and are usually compared with specific distances, such as the bottleneck distance. To incorporate PDs in a learning pipeline, several kernels have been proposed for PDs with a strong emphasis on the stability of the RKHS distance w.r.t. perturbations of the PDs. In this article, we use the Sliced Wasserstein approximation of the Wasserstein distance to define a new kernel for PDs, which is not only provably stable but also provably discriminative w.r.t. the Wasserstein distance W^1_∞ between PDs. We also demonstrate its practicality, by developing an approximation technique to reduce kernel computation time, and show that our proposal compares favorably to existing kernels for PDs on several benchmarks.

****

Multiple Clustering Views from Multiple Uncertain Experts

Yale Chang, Junxiang Chen, Michael H. Cho, Peter J. Castaldi, Edwin K. Silverman, Jennifer G. Dy

Expert input can improve clustering performance. In today’s collaborative environment, the availability of crowdsourced multiple expert input is becoming common. Given multiple experts’ inputs, most existing approaches can only discover one clustering structure. However, data is multi-faced by nature and can be clustered in different ways (also known as views). In an exploratory analysis problem where ground truth is not known, different experts may have diverse views on how to cluster data. In this paper, we address the problem on how to automatically discover multiple ways to cluster data given potentially diverse inputs from multiple uncertain experts. We propose a novel Bayesian probabilistic model that automatically learns the multiple expert views and the clustering structure associated with each view. The benefits of learning the experts’ views include 1) enabling the discovery of multiple diverse clustering structures, and 2) improving the quality of clustering solution in each view by assigning higher weights to experts with higher confidence. In our approach, the expert views, multiple clustering structures and expert confidences are jointly learned via variational inference. Experimental results on synthetic datasets, benchmark datasets and a real-world disease subtyping problem show that our proposed approach outperforms competing baselines, including meta clustering, semi-supervised clustering, semi-crowdsourced clustering and consensus clustering.

****

Uncertainty Assessment and False Discovery Rate Control in High-Dimensional Granger Causal Inference

Aditya Chaudhry, Pan Xu, Quanquan Gu

Causal inference among high-dimensional time series data proves an important research problem in many fields. While in the classical regime one often establishes causality among time series via a concept known as “Granger causality,” existing approaches for Granger causal inference in high-dimensional data lack the means to characterize the uncertainty associated with Granger causality estimates (e.g., p-values and confidence intervals). We make two contributions in this work. First, we introduce a novel asymptotically unbiased Granger causality estimator with corresponding test statistics and confidence intervals to allow, for the first time, uncertainty characterization in high-dimensional Granger causal inference. Second, we introduce a novel method for false discovery rate control that achieves higher power in multiple testing than existing techniques and that can cope with dependent test statistics and dependent observations. We corroborate our theoretical results with experiments on both synthetic data and real-world climatological data.

****

Active Heteroscedastic Regression
Kamalika Chaudhuri, Prateek Jain, Nagarajan Natarajan

An active learner is given a model class Θ, a large sample of unlabeled data drawn from an underlying distribution and access to a labeling oracle that can provide a label for any of the unlabeled instances. The goal of the learner is to find a model $\theta \in \Theta$ that fits the data to a given accuracy while making as few label queries to the oracle as possible. In this work, we consider a theoretical analysis of the label requirement of active learning for regression under a heteroscedastic noise model, where the noise depends on the instance. We provide bounds on the convergence rates of active and passive learning for heteroscedastic regression. Our results illustrate that just like in binary classification, some partial knowledge of the nature of the noise can lead to significant gains in the label requirement of active learning.

****

Combining Model-Based and Model-Free Updates for Trajectory-Centric Reinforcement Learning
Yevgen Chebotar, Karol Hausman, Marvin Zhang, Gaurav Sukhatme, Stefan Schaal, Sergey Levine

Reinforcement learning algorithms for real-world robotic applications must be able to handle complex, unknown dynamical systems while maintaining data-efficient learning. These requirements are handled well by model-free and model-based RL approaches, respectively. In this work, we aim to combine the advantages of these approaches. By focusing on time-varying linear-Gaussian policies, we enable a model-based algorithm based on the linear-quadratic regulator that can be integrated into the model-free framework of path integral policy improvement. We can further combine our method with guided policy search to train arbitrary parameterized policies such as deep neural networks. Our simulation and real-world experiments demonstrate that this method can solve challenging manipulation tasks with comparable or better performance than model-free methods while maintaining the sample efficiency of model-based methods.

****

Robust Structured Estimation with Single-Index Models
Sheng Chen, Arindam Banerjee

In this paper, we investigate general single-index models (SIMs) in high dimensions. Based on U-statistics, we propose two types of robust estimators for the recovery of model parameters, which can be viewed as generalizations of several existing algorithms for one-bit compressed sensing (1-bit CS). With minimal assumption on noise, the statistical guarantees are established for the generalized estimators under suitable conditions, which allow general structures of underlying parameter. Moreover, the proposed estimator is novelly instantiated for SIMs with monotone transfer function, and the obtained estimator can better leverage the monotonicity. Experimental results are provided to support our theoretical analyses.

****

Adaptive Multiple-Arm Identification

Jiecao Chen, Xi Chen, Qin Zhang, Yuan Zhou

We study the problem of selecting K arms with the highest expected rewards in a stochastic n-armed bandit game. This problem has a wide range of applications, e.g., A/B testing, crowdsourcing, simulation optimization. Our goal is to develop a PAC algorithm, which, with probability at least $1-\delta$, identifies a set of K arms with the aggregate regret at most ϵ. The notion of aggregate regret for multiple-arm identification was first introduced in Zhou et. al. (2014), which is defined as the difference of the averaged expected rewards between the selected set of arms and the best K arms. In contrast to Zhou et. al. (2014) that only provides instance-independent sample complexity, we introduce a new hardness parameter for characterizing the difficulty of any given instance. We further develop two algorithms and establish the corresponding sample complexity in terms of this hardness parameter. The derived sample complexity can be significantly smaller than state-of-the-art results for a large class of instances and matches the instance-independent lower bound up to a $\log(\epsilon^{-1})$ factor in the worst case. We also prove a lower bound result showing that the extra $\log(\epsilon^{-1})$ is necessary for instance-dependent algorithms using the introduced hardness parameter.

****

Dueling Bandits with Weak Regret

Bangrui Chen, Peter I. Frazier

We consider online content recommendation with implicit feedback through pairwise comparisons, formalized as the so-called dueling bandit problem. We study the dueling bandit problem in the Condorcet winner setting, and consider two notions of regret: the more well-studied strong regret, which is 0 only when both arms pulled are the Condorcet winner; and the less well-studied weak regret, which is 0 if either arm pulled is the Condorcet winner. We propose a new algorithm for this problem, Winner Stays (WS), with variations for each kind of regret: WS for weak regret (WS-W) has expected cumulative weak regret that is $O(N^2)$, and $O(N\log(N))$ if arms have a total order; WS for strong regret (WS-S) has expected cumulative strong regret of $O(N^2 + N \log(T))$, and $O(N\log(N)+N\log(T))$ if arms have a total order. WS-W is the first dueling bandit algorithm with weak regret that is constant in time. WS is simple to compute, even for problems with many arms, and we demonstrate through numerical experiments on simulated and real data that WS has significantly smaller regret than existing algorithms in both the weak- and strong-regret settings.

****

Strong NP-Hardness for Sparse Optimization with Concave Penalty Functions

Yichen Chen, Dongdong Ge, Mengdi Wang, Zizhuo Wang, Yinyu Ye, Hao Yin

Consider the regularized sparse minimization problem, which involves empirical sums of loss functions for n data points (each of dimension d) and a nonconvex sparsity penalty. We prove that finding an $\mathcal{O}(n^{c_1}d^{c_2})$-optimal solution to the regularized sparse optimization problem is strongly NP-hard for any $c_1, c_2\in [0,1)$ such that $c_1+c_2<1$. The result applies to a broad class of loss functions and sparse penalty functions. It suggests that one cannot even approximately solve the sparse optimization problem in polynomial time, unless P $=$ NP.

****

Learning to Learn without Gradient Descent by Gradient Descent

Yutian Chen, Matthew W. Hoffman, Sergio Gómez Colmenarejo, Misha Denil, Timothy P. Lillicrap, Matt Botvinick, Nando Freitas

We learn recurrent neural network optimizers trained on simple synthetic functions by gradient descent. We show that these learned optimizers exhibit a remarkable degree of transfer in that they can be used to efficiently optimize a broad range of derivative-free black-box functions, including Gaussian process bandits, simple control objectives, global optimization benchmarks and hyper-parameter tuning tasks. Up to the training horizon, the learned optimizers learn to trade-off exploration and exploitation, and compare favourably with heavily engineered Bayesian optimization packages for hyper-parameter tuning.

****

Identification and Model Testing in Linear Structural Equation Models using Auxiliary Variables

Bryant Chen, Daniel Kumor, Elias Bareinboim

We developed a novel approach to identification and model testing in linear structural equation models (SEMs) based on auxiliary variables (AVs), which generalizes a widely-used family of methods known as instrumental variables. The identification problem is concerned with the conditions under which causal parameters can be uniquely estimated from an observational, non-causal covariance matrix. In this paper, we provide an algorithm for the identification of causal parameters in linear structural models that subsumes previous state-of-the-art methods. In other words, our algorithm identifies strictly more coefficients and models than methods previously known in the literature. Our algorithm builds on a graph-theoretic characterization of conditional independence relations between auxiliary and model variables, which is developed in this paper. Further, we leverage this new characterization for allowing identification when limited experimental data or new substantive knowledge about the domain is available. Lastly, we develop a new procedure for model testing using AVs.

****

Toward Efficient and Accurate Covariance Matrix Estimation on Compressed Data

Xixian Chen, Michael R. Lyu, Irwin King

Estimating covariance matrices is a fundamental technique in various domains, most notably in machine learning and signal processing. To tackle the challenges of extensive communication costs, large storage capacity requirements, and high processing time complexity when handling massive high-dimensional and distributed data, we propose an efficient and accurate covariance matrix estimation method via data compression. In contrast to previous data-oblivious compression schemes, we leverage a data-aware weighted sampling method to construct low-dimensional data for such estimation. We rigorously prove that our proposed estimator is unbiased and requires smaller data to achieve the same accuracy with specially designed sampling distributions. Besides, we depict that the computational procedures in our algorithm are efficient. All achievements imply an improved tradeoff between the estimation accuracy and computational costs. Finally, the extensive experiments on synthetic and real-world datasets validate the superior property of our method and illustrate that it significantly outperforms the state-of-the-art algorithms.

****

Online Partial Least Square Optimization: Dropping Convexity for Better Efficiency and Scalability

Zhehui Chen, Lin F. Yang, Chris Junchi Li, Tuo Zhao

Multiview representation learning is popular for latent factor analysis. Many existing approaches formulate the multiview representation learning as convex optimization problems, where global optima can be obtained by certain algorithms in polynomial time. However, many evidences have corroborated that heuristic nonconvex approaches also have good empirical computational performance and convergence to the global optima, although there is a lack of theoretical justification. Such a gap between theory and practice motivates us to study a nonconvex formulation for multiview representation learning, which can be efficiently solved by a simple stochastic gradient descent method. By analyzing the dynamics of the algorithm based on diffusion processes, we establish a global rate of convergence to the global optima. Numerical experiments are provided to support our theory.

****

Learning to Aggregate Ordinal Labels by Maximizing Separating Width

Guangyong Chen, Shengyu Zhang, Di Lin, Hui Huang, Pheng Ann Heng

While crowdsourcing has been a cost and time efficient method to label massive samples, one critical issue is quality control, for which the key challenge is to infer the ground truth from noisy or even adversarial data by various users. A large class of crowdsourcing problems, such as those involving age, grade, level, or stage, have an ordinal structure in their labels. Based on a technique of sampling estimated label from the posterior distribution, we define a novel separating width among the labeled observations to characterize the quality of sample

d labels, and develop an efficient algorithm to optimize it through solving multiple linear decision boundaries and adjusting prior distributions. Our algorithm is empirically evaluated on several real world datasets, and demonstrates its supremacy over state-of-the-art methods.
****

Nearly Optimal Robust Matrix Completion
Yeshwanth Cherapanamjeri, Kartik Gupta, Prateek Jain
In this paper, we consider the problem of Robust Matrix Completion (RMC) where the goal is to recover a low-rank matrix by observing a small number of its entries out of which a few can be arbitrarily corrupted. We propose a simple projected gradient descent-based method to estimate the low-rank matrix that alternately performs a projected gradient descent step and cleans up a few of the corrupted entries using hard-thresholding. Our algorithm solves RMC using nearly optimal number of observations while tolerating a nearly optimal number of corruptions. Our result also implies significant improvement over the existing time complexity bounds for the low-rank matrix completion problem. Finally, an application of our result to the robust PCA problem (low-rank+sparse matrix separation) leads to nearly linear time (in matrix dimensions) algorithm for the same; existing state-of-the-art methods require quadratic time. Our empirical results corroborate our theoretical results and show that even for moderate sized problems, our method for robust PCA is an order of magnitude faster than the existing methods.
****

Algorithms for ℓ_p Low-Rank Approximation
Flavio Chierichetti, Sreenivas Gollapudi, Ravi Kumar, Silvio Lattanzi, Rina Panigrahy, David P. Woodruff
We consider the problem of approximating a given matrix by a low-rank matrix so as to minimize the entrywise ℓ_p-approximation error, for any $p \geq 1$; the case $p = 2$ is the classical SVD problem. We obtain the first provably good approximation algorithms for this robust version of low-rank approximation that work for every value of p. Our algorithms are simple, easy to implement, work well in practice, and illustrate interesting tradeoffs between the approximation quality, the running time, and the rank of the approximating matrix.
****

MEC: Memory-efficient Convolution for Deep Neural Network
Minsik Cho, Daniel Brand
Convolution is a critical component in modern deep neural networks, thus several algorithms for convolution have been developed. Direct convolution is simple but suffers from poor performance. As an alternative, multiple indirect methods have been proposed including im2col-based convolution, FFT-based convolution, or Winograd-based algorithm. However, all these indirect methods have high memory overhead, which creates performance degradation and offers a poor trade-off between performance and memory consumption. In this work, we propose a memory-efficient convolution or MEC with compact lowering, which reduces memory overhead substantially and accelerates convolution process. MEC lowers the input matrix in a simple yet efficient/compact way (i.e., much less memory overhead), and then executes multiple small matrix multiplications in parallel to get convolution completed. Additionally, the reduced memory footprint improves memory sub-system efficiency, improving performance. Our experimental results show that MEC reduces memory consumption significantly with good speedup on both mobile and server platforms, compared with other indirect convolution algorithms.
****

On Relaxing Determinism in Arithmetic Circuits
Arthur Choi, Adnan Darwiche
The past decade has seen a significant interest in learning tractable probabilistic representations. Arithmetic circuits (ACs) were among the first proposed tractable representations, with some subsequent representations being instances of ACs with weaker or stronger properties. In this paper, we provide a formal basis under which variants on ACs can be compared, and where the precise roles and semantics of their various properties can be made more transparent. This allows us to place some recent developments on ACs in a clearer perspective and to also d

erive new results for ACs. This includes an exponential separation between ACs with and without determinism; completeness and incompleteness results; and tractability results (or lack thereof) when computing most probable explanations (MPEs).

****

Improving Stochastic Policy Gradients in Continuous Control with Deep Reinforcement Learning using the Beta Distribution

Po-Wei Chou, Daniel Maturana, Sebastian Scherer

Recently, reinforcement learning with deep neural networks has achieved great success in challenging continuous control problems such as 3D locomotion and robotic manipulation. However, in real-world control problems, the actions one can take are bounded by physical constraints, which introduces a bias when the standard Gaussian distribution is used as the stochastic policy. In this work, we propose to use the Beta distribution as an alternative and analyze the bias and variance of the policy gradients of both policies. We show that the Beta policy is bias-free and provides significantly faster convergence and higher scores over the Gaussian policy when both are used with trust region policy optimization (TRPO) and actor critic with experience replay (ACER), the state-of-the-art on- and off-policy stochastic methods respectively, on OpenAI Gym’s and MuJoCo’s continuous control environments.

****

On Kernelized Multi-armed Bandits

Sayak Ray Chowdhury, Aditya Gopalan

We consider the stochastic bandit problem with a continuous set of arms, with the expected reward function over the arms assumed to be fixed but unknown. We provide two new Gaussian process-based algorithms for continuous bandit optimization – Improved GP-UCB (IGP-UCB) and GP-Thomson sampling (GP-TS), and derive corresponding regret bounds. Specifically, the bounds hold when the expected reward function belongs to the reproducing kernel Hilbert space (RKHS) that naturally corresponds to a Gaussian process kernel used as input by the algorithms. Along the way, we derive a new self-normalized concentration inequality for vector-valued martingales of arbitrary, possibly infinite, dimension. Finally, experimental evaluation and comparisons to existing algorithms on synthetic and real-world environments are carried out that highlight the favourable gains of the proposed strategies in many cases.

****

Parseval Networks: Improving Robustness to Adversarial Examples

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, Nicolas Usunier

We introduce Parseval networks, a form of deep neural networks in which the Lipschitz constant of linear, convolutional and aggregation layers is constrained to be smaller than 1. Parseval networks are empirically and theoretically motivated by an analysis of the robustness of the predictions made by deep neural networks when their input is subject to an adversarial perturbation. The most important feature of Parseval networks is to maintain weight matrices of linear and convolutional layers to be (approximately) Parseval tight frames, which are extensions of orthogonal matrices to non-square matrices. We describe how these constraints can be maintained efficiently during SGD. We show that Parseval networks match the state-of-the-art regarding accuracy on CIFAR-10/100 and Street View House Numbers (SVHN), while being more robust than their vanilla counterpart against adversarial examples. Incidentally, Parseval networks also tend to train faster and make a better usage of the full capacity of the networks.

****

Deep Latent Dirichlet Allocation with Topic-Layer-Adaptive Stochastic Gradient Riemannian MCMC

Yulai Cong, Bo Chen, Hongwei Liu, Mingyuan Zhou

It is challenging to develop stochastic gradient based scalable inference for deep discrete latent variable models (LVMs), due to the difficulties in not only computing the gradients, but also adapting the step sizes to different latent factors and hidden layers. For the Poisson gamma belief network (PGBN), a recently proposed deep discrete LVM, we derive an alternative representation that is refe rred to as deep latent Dirichlet allocation (DLDA). Exploiting data augmentation and marginalization techniques, we derive a block-diagonal Fisher information matrix and its inverse for the simplex-constrained global model parameters of DLDA. Exploiting that Fisher information matrix with stochastic gradient MCMC, we present topic-layer-adaptive stochastic gradient Riemannian (TLASGR) MCMC that jointly learns simplex-constrained global parameters across all layers and topics, with topic and layer specific learning rates. State-of-the-art results are demonstrated on big data sets.

****

AdaNet: Adaptive Structural Learning of Artificial Neural Networks

Corinna Cortes, Xavier Gonzalvo, Vitaly Kuznetsov, Mehryar Mohri, Scott Yang

We present a new framework for analyzing and learning artificial neural networks. Our approach simultaneously and adaptively learns both the structure of the network as well as its weights. The methodology is based upon and accompanied by strong data-dependent theoretical learning guarantees, so that the final network architecture provably adapts to the complexity of any given problem.

****

Random Feature Expansions for Deep Gaussian Processes

Kurt Cutajar, Edwin V. Bonilla, Pietro Michiardi, Maurizio Filippone

The composition of multiple Gaussian Processes as a Deep Gaussian Process DGP enables a deep probabilistic nonparametric approach to flexibly tackle complex machine learning problems with sound quantification of uncertainty. Existing inference approaches for DGP models have limited scalability and are notoriously cumbersome to construct. In this work we introduce a novel formulation of DGPs based on random feature expansions that we train using stochastic variational inference. This yields a practical learning framework which significantly advances the state-of-the-art in inference for DGPs, and enables accurate quantification of uncertainty. We extensively showcase the scalability and performance of our proposal on several datasets with up to 8 million observations, and various DGP architectures with up to 30 hidden layers.

****

Soft-DTW: a Differentiable Loss Function for Time-Series

Marco Cuturi, Mathieu Blondel

We propose in this paper a differentiable learning loss between time series, building upon the celebrated dynamic time warping (DTW) discrepancy. Unlike the Euclidean distance, DTW can compare time series of variable size and is robust to shifts or dilatations across the time dimension. To compute DTW, one typically solves a minimal-cost alignment problem between two time series using dynamic programming. Our work takes advantage of a smoothed formulation of DTW, called soft-DTW, that computes the soft-minimum of all alignment costs. We show in this paper that soft-DTW is a differentiable loss function, and that both its value and gradient can be computed with quadratic time/space complexity (DTW has quadratic time but linear space complexity). We show that this regularization is particularly well suited to average and cluster time series under the DTW geometry, a task for which our proposal significantly outperforms existing baselines (Petitjean et al., 2011). Next, we propose to tune the parameters of a machine that outputs time series by minimizing its fit with ground-truth labels in a soft-DTW sense. Source code is available at https://github.com/mblondel/soft-dtw

****

Understanding Synthetic Gradients and Decoupled Neural Interfaces

Wojciech Marian Czarnecki, Grzegorz ■wirszcz, Max Jaderberg, Simon Osindero, Oriol Vinyals, Koray Kavukcuoglu

When training neural networks, the use of Synthetic Gradients (SG) allows layers or modules to be trained without update locking – without waiting for a true error gradient to be backpropagated – resulting in Decoupled Neural Interfaces (DNIs). This unlocked ability of being able to update parts of a neural network asynchronously and with only local information was demonstrated to work empirically in Jaderberg et al (2016). However, there has been very little demonstration of what changes DNIs and SGs impose from a functional, representational, and learning dynamics point of view. In this paper, we study DNIs through the use of synt hetic gradients on feed-forward networks to better understand their behaviour and elucidate their effect on optimisation. We show that the incorporation of SGs does not affect the representational strength of the learning system for a neural network, and prove the convergence of the learning system for linear and deep linear models. On practical problems we investigate the mechanism by which synthetic gradient estimators approximate the true loss, and, surprisingly, how that leads to drastically different layer-wise representations. Finally, we also expose the relationship of using synthetic gradients to other error approximation techniques and find a unifying language for discussion and comparison.

****

Stochastic Generative Hashing

Bo Dai, Ruiqi Guo, Sanjiv Kumar, Niao He, Le Song

Learning-based binary hashing has become a powerful paradigm for fast search and retrieval in massive databases. However, due to the requirement of discrete outputs for the hash functions, learning such functions is known to be very challenging. In addition, the objective functions adopted by existing hashing techniques are mostly chosen heuristically. In this paper, we propose a novel generative approach to learn hash functions through Minimum Description Length principle such that the learned hash codes maximally compress the dataset and can also be used to regenerate the inputs. We also develop an efficient learning algorithm based on the stochastic distributional gradient, which avoids the notorious difficulty caused by binary output constraints, to jointly optimize the parameters of the hash function and the associated generative model. Extensive experiments on a variety of large-scale datasets show that the proposed method achieves better retrieval results than the existing state-of-the-art methods.

****

Logarithmic Time One-Against-Some

Hal Daumé III, Nikos Karampatziakis, John Langford, Paul Mineiro

We create a new online reduction of multiclass classification to binary classification for which training and prediction time scale logarithmically with the number of classes. We show that several simple techniques give rise to an algorithm which is superior to previous logarithmic time classification approaches while competing with one-against-all in space. The core construction is based on using a tree to select a small subset of labels with high recall, which are then scored using a one-against-some structure with high precision.

****

Language Modeling with Gated Convolutional Networks

Yann N. Dauphin, Angela Fan, Michael Auli, David Grangier

The pre-dominant approach to language modeling to date is based on recurrent neural networks. Their success on this task is often linked to their ability to capture unbounded context. In this paper we develop a finite context approach through stacked convolutions, which can be more efficient since they allow parallelization over sequential tokens. We propose a novel simplified gating mechanism that outperforms Oord et al. (2016) and investigate the impact of key architectural decisions. The proposed approach achieves state-of-the-art on the WikiText-103 benchmark, even though it features long-term dependencies, as well as competitive results on the Google Billion Words benchmark. Our model reduces the latency to score a sentence by an order of magnitude compared to a recurrent baseline. To our knowledge, this is the first time a non-recurrent approach is competitive with strong recurrent models on these large scale language tasks.

****

An Infinite Hidden Markov Model With Similarity-Biased Transitions

Colin Reimer Dawson, Chaofan Huang, Clayton T. Morrison

We describe a generalization of the Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM) which is able to encode prior information that state transitions are more likely between “nearby” states. This is accomplished by defining a similarity function on the state space and scaling transition probabilities by pairwise similarities, thereby inducing correlations among the transition distributions. We present an augmented data representation of the model as a Markov Jump Process in which: (1) some jump attempts fail, and (2) the probability of succes

s is proportional to the similarity between the source and destination states. This augmentation restores conditional conjugacy and admits a simple Gibbs sampler. We evaluate the model and inference method on a speaker diarization task and a “harmonic parsing” task using four-part chorale data, as well as on several synthetic datasets, achieving favorable comparisons to existing models.
****
Distributed Batch Gaussian Process Optimization
Erik A. Daxberger, Bryan Kian Hsiang Low
This paper presents a novel distributed batch Gaussian process upper confidence bound (DB-GP-UCB) algorithm for performing batch Bayesian optimization (BO) of highly complex, costly-to-evaluate black-box objective functions. In contrast to existing batch BO algorithms, DB-GP-UCB can jointly optimize a batch of inputs (as opposed to selecting the inputs of a batch one at a time) while still preserving scalability in the batch size. To realize this, we generalize GP-UCB to a new batch variant amenable to a Markov approximation, which can then be naturally formulated as a multi-agent distributed constraint optimization problem in order to fully exploit the efficiency of its state-of-the-art solvers for achieving linear time in the batch size. Our DB-GP-UCB algorithm offers practitioners the flexibility to trade off between the approximation quality and time efficiency by varying the Markov order. We provide a theoretical guarantee for the convergence rate of DB-GP-UCB via bounds on its cumulative regret. Empirical evaluation on synthetic benchmark objective functions and a real-world optimization problem shows that DB-GP-UCB outperforms the state-of-the-art batch BO algorithms.
****
Consistency Analysis for Binary Classification Revisited
Krzysztof Dembczy■ski, Wojciech Kot■owski, Oluwasanmi Koyejo, Nagarajan Natarajan
Statistical learning theory is at an inflection point enabled by recent advances in understanding and optimizing a wide range of metrics. Of particular interest are non-decomposable metrics such as the F-measure and the Jaccard measure which cannot be represented as a simple average over examples. Non-decomposability is the primary source of difficulty in theoretical analysis, and interestingly has led to two distinct settings and notions of consistency. In this manuscript we analyze both settings, from statistical and algorithmic points of view, to explore the connections and to highlight differences between them for a wide range of metrics. The analysis complements previous results on this topic, clarifies common confusions around both settings, and provides guidance to the theory and practice of binary classification with complex metrics.
****
iSurvive: An Interpretable, Event-time Prediction Model for mHealth
Walter H. Dempsey, Alexander Moreno, Christy K. Scott, Michael L. Dennis, David H. Gustafson, Susan A. Murphy, James M. Rehg
An important mobile health (mHealth) task is the use of multimodal data, such as sensor streams and self-report, to construct interpretable time-to-event predictions of, for example, lapse to alcohol or illicit drug use. Interpretability of the prediction model is important for acceptance and adoption by domain scientists, enabling model outputs and parameters to inform theory and guide intervention design. Temporal latent state models are therefore attractive, and so we adopt the continuous time hidden Markov model (CT-HMM) due to its ability to describe irregular arrival times of event data. Standard CT-HMMs, however, are not specialized for predicting the time to a future event, the key variable for mHealth interventions. Also, standard emission models lack a sufficiently rich structure to describe multimodal data and incorporate domain knowledge. We present iSurvive, an extension of classical survival analysis to a CT-HMM. We present a parameter learning method for GLM emissions and survival model fitting, and present promising results on both synthetic data and an mHealth drug use dataset.
****
Image-to-Markup Generation with Coarse-to-Fine Attention
Yuntian Deng, Anssi Kanervisto, Jeffrey Ling, Alexander M. Rush
We present a neural encoder-decoder model to convert images into presentational

markup based on a scalable coarse-to-fine attention mechanism. Our method is evaluated in the context of image-to-LaTeX generation, and we introduce a new dataset of real-world rendered mathematical expressions paired with LaTeX markup. We show that unlike neural OCR techniques using CTC-based models, attention-based approaches can tackle this non-standard OCR task. Our approach outperforms classical mathematical OCR systems by a large margin on in-domain rendered data, and, with pretraining, also performs well on out-of-domain handwritten data. To reduce the inference complexity associated with the attention-based approaches, we introduce a new coarse-to-fine attention layer that selects a support region before applying attention.

****

RobustFill: Neural Program Learning under Noisy I/O

Jacob Devlin, Jonathan Uesato, Surya Bhupatiraju, Rishabh Singh, Abdel-rahman Mohamed, Pushmeet Kohli

The problem of automatically generating a computer program from some specification has been studied since the early days of AI. Recently, two competing approaches for `automatic program learning' have received significant attention: (1) `neural program synthesis', where a neural network is conditioned on input/output (I/O) examples and learns to generate a program, and (2) `neural program induction', where a neural network generates new outputs directly using a latent program representation. Here, for the first time, we directly compare both approaches on a large-scale, real-world learning task and we additionally contrast to rule-based program synthesis, which uses hand-crafted semantics to guide the program generation. Our neural models use a modified attention RNN to allow encoding of variable-sized sets of I/O pairs, which achieve 92\% accuracy on a real-world test set, compared to the 34\% accuracy of the previous best neural synthesis approach. The synthesis model also outperforms a comparable induction model on this task, but we more importantly demonstrate that the strength of each approach is highly dependent on the evaluation metric and end-user application. Finally, we show that we can train our neural models to remain very robust to the type of noise expected in real-world data (e.g., typos), while a highly-engineered rule-based system fails entirely.

****

Being Robust (in High Dimensions) Can Be Practical

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, Alistair Stewart

Robust estimation is much more challenging in high-dimensions than it is in one-dimension: Most techniques either lead to intractable optimization problems or estimators that can tolerate only a tiny fraction of errors. Recent work in theoretical computer science has shown that, in appropriate distributional models, it is possible to robustly estimate the mean and covariance with polynomial time algorithms that can tolerate a constant fraction of corruptions, independent of the dimension. However, the sample and time complexity of these algorithms is prohibitively large for high-dimensional applications. In this work, we address both of these issues by establishing sample complexity bounds that are optimal, up to logarithmic factors, as well as giving various refinements that allow the algorithms to tolerate a much larger fraction of corruptions. Finally, we show on both synthetic and real data that our algorithms have state-of-the-art performance and suddenly make high-dimensional robust estimation a realistic possibility.

****

Probabilistic Path Hamiltonian Monte Carlo

Vu Dinh, Arman Bilge, Cheng Zhang, Frederick A. Matsen IV

Hamiltonian Monte Carlo (HMC) is an efficient and effective means of sampling posterior distributions on Euclidean space, which has been extended to manifolds with boundary. However, some applications require an extension to more general spaces. For example, phylogenetic (evolutionary) trees are defined in terms of both a discrete graph and associated continuous parameters; although one can represent these aspects using a single connected space, this rather complex space is not suitable for existing HMC algorithms. In this paper, we develop Probabilistic Path HMC (PPHMC) as a first step to sampling distributions on spaces with intri cate combinatorial structure. We define PPHMC on orthant complexes, show that the resulting Markov chain is ergodic, and provide a promising implementation for the case of phylogenetic trees in open-source software. We also show that a surrogate function to ease the transition across a boundary on which the log-posterior has discontinuous derivatives can greatly improve efficiency.

****

Sharp Minima Can Generalize For Deep Nets

Laurent Dinh, Razvan Pascanu, Samy Bengio, Yoshua Bengio

Despite their overwhelming capacity to overfit, deep learning architectures tend to generalize relatively well to unseen data, allowing them to be deployed in practice. However, explaining why this is the case is still an open area of research. One standing hypothesis that is gaining popularity, e.g.\ Hochreiter \& Schmidhuber (1997); Keskar et al.\ (2017), is that the flatness of minima of the loss function found by stochastic gradient based methods results in good generalization. This paper argues that most notions of flatness are problematic for deep models and can not be directly applied to explain generalization. Specifically, when focusing on deep networks with rectifier units, we can exploit the particular geometry of parameter space induced by the inherent symmetries that these architectures exhibit to build equivalent models corresponding to arbitrarily sharper minima. Or, depending on the definition of flatness, it is the same for any given minimum. Furthermore, if we allow to reparametrize a function, the geometry of its parameters can change drastically without affecting its generalization properties.

****

A Divergence Bound for Hybrids of MCMC and Variational Inference and an Application to Langevin Dynamics and SGVI

Justin Domke

Two popular classes of methods for approximate inference are Markov chain Monte Carlo (MCMC) and variational inference. MCMC tends to be accurate if run for a long enough time, while variational inference tends to give better approximations at shorter time horizons. However, the amount of time needed for MCMC to exceed the performance of variational methods can be quite high, motivating more fine-grained tradeoffs. This paper derives a distribution over variational parameters, designed to minimize a bound on the divergence between the resulting marginal distribution and the target, and gives an example of how to sample from this distribution in a way that interpolates between the behavior of existing methods based on Langevin dynamics and stochastic gradient variational inference (SGVI).

****

Dance Dance Convolution

Chris Donahue, Zachary C. Lipton, Julian McAuley

Dance Dance Revolution (DDR) is a popular rhythm-based video game. Players perform steps on a dance platform in synchronization with music as directed by on-screen step charts. While many step charts are available in standardized packs, players may grow tired of existing charts, or wish to dance to a song for which no chart exists. We introduce the task of learning to choreograph. Given a raw audio track, the goal is to produce a new step chart. This task decomposes naturally into two subtasks: deciding when to place steps and deciding which steps to select. For the step placement task, we combine recurrent and convolutional neural networks to ingest spectrograms of low-level audio features to predict steps, conditioned on chart difficulty. For step selection, we present a conditional LSTM generative model that substantially outperforms n-gram and fixed-window approaches.

****

Stochastic Variance Reduction Methods for Policy Evaluation

Simon S. Du, Jianshu Chen, Lihong Li, Lin Xiao, Dengyong Zhou

Policy evaluation is concerned with estimating the value function that predicts long-term values of states under a given policy. It is a crucial step in many reinforcement-learning algorithms. In this paper, we focus on policy evaluation with linear function approximation over a fixed dataset. We first transform the empirical policy evaluation problem into a (quadratic) convex-concave saddle-point

problem, and then present a primal-dual batch gradient method, as well as two stochastic variance reduction methods for solving the problem. These algorithms scale linearly in both sample size and feature dimension. Moreover, they achieve linear convergence even when the saddle-point problem has only strong concavity in the dual variables but no strong convexity in the primal variables. Numerical experiments on benchmark problems demonstrate the effectiveness of our methods.
****

Rule-Enhanced Penalized Regression by Column Generation using Rectangular Maximum Agreement

Jonathan Eckstein, Noam Goldberg, Ai Kagawa

We describe a learning procedure enhancing L1-penalized regression by adding dynamically generated rules describing multidimensional “box” sets. Our rule-adding procedure is based on the classical column generation method for high-dimensional linear programming. The pricing problem for our column generation procedure reduces to the NP-hard rectangular maximum agreement (RMA) problem of finding a box that best discriminates between two weighted datasets. We solve this problem exactly using a parallel branch-and-bound procedure. The resulting rule-enhanced regression procedure is computation-intensive, but has promising prediction performance.
****

Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, Karen Simonyan

Generative models in vision have seen rapid progress due to algorithmic improvements and the availability of high-quality image datasets. In this paper, we offer contributions in both these areas to enable similar progress in audio modeling. First, we detail a powerful new WaveNet-style autoencoder model that conditions an autoregressive decoder on temporal codes learned from the raw audio waveform. Second, we introduce NSynth, a large-scale and high-quality dataset of musical notes that is an order of magnitude larger than comparable public datasets. Using NSynth, we demonstrate improved qualitative and quantitative performance of the WaveNet autoencoder over a well-tuned spectral autoencoder baseline. Finally, we show that the model learns a manifold of embeddings that allows for morphing between instruments, meaningfully interpolating in timbre to create new types of sounds that are realistic and expressive.
****

Statistical Inference for Incomplete Ranking Data: The Case of Rank-Dependent Coarsening

Mohsen Ahmadi Fahandar, Eyke Hüllermeier, Inés Couso

We consider the problem of statistical inference for ranking data, specifically rank aggregation, under the assumption that samples are incomplete in the sense of not comprising all choice alternatives. In contrast to most existing methods, we explicitly model the process of turning a full ranking into an incomplete one, which we call the coarsening process. To this end, we propose the concept of rank-dependent coarsening, which assumes that incomplete rankings are produced by projecting a full ranking to a random subset of ranks. For a concrete instantiation of our model, in which full rankings are drawn from a Plackett-Luce distribution and observations take the form of pairwise preferences, we study the performance of various rank aggregation methods. In addition to predictive accuracy in the finite sample setting, we address the theoretical question of consistency, by which we mean the ability to recover a target ranking when the sample size goes to infinity, despite a potential bias in the observations caused by the (unknown) coarsening.
****

Maximum Selection and Ranking under Noisy Comparisons

Moein Falahatgar, Alon Orlitsky, Venkatadheeraj Pichapati, Ananda Theertha Suresh

We consider (ϵ,δ)-PAC maximum-selection and ranking using pairwise comparisons for general probabilistic models whose comparison probabilities satisfy strong stochastic transitivity and stochastic triangle inequality. Modifyin g the popular knockout tournament, we propose a simple maximum-selection algorithm that uses $\mathcal{O}\left(\frac{n}{\epsilon^2} \left(1+\log \frac1{\delta}\right)\right)$ comparisons, optimal up to a constant factor. We then derive a general framework that uses noisy binary search to speed up many ranking algorithms, and combine it with merge sort to obtain a ranking algorithm that uses $\mathcal{O}\left(\frac n{\epsilon^2}\log n(\log \log n)^3\right)$ comparisons for $\delta=\frac1n$, optimal up to a $(\log \log n)^3$ factor.

****

Fake News Mitigation via Point Process Based Intervention

Mehrdad Farajtabar, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, Hongyuan Zha

We propose the first multistage intervention framework that tackles fake news in social networks by combining reinforcement learning with a point process network activity model. The spread of fake news and mitigation events within the network is modeled by a multivariate Hawkes process with additional exogenous control terms. By choosing a feature representation of states, defining mitigation actions and constructing reward functions to measure the effectiveness of mitigation activities, we map the problem of fake news mitigation into the reinforcement learning framework. We develop a policy iteration method unique to the multivariate networked point process, with the goal of optimizing the actions for maximal reward under budget constraints. Our method shows promising performance in real-time intervention experiments on a Twitter network to mitigate a surrogate fake news campaign, and outperforms alternatives on synthetic datasets.

****

Regret Minimization in Behaviorally-Constrained Zero-Sum Games

Gabriele Farina, Christian Kroer, Tuomas Sandholm

No-regret learning has emerged as a powerful tool for solving extensive-form games. This was facilitated by the counterfactual-regret minimization (CFR) framework, which relies on the instantiation of regret minimizers for simplexes at each information set of the game. We use an instantiation of the CFR framework to develop algorithms for solving behaviorally-constrained (and, as a special case, perturbed in the Selten sense) extensive-form games, which allows us to compute approximate Nash equilibrium refinements. Nash equilibrium refinements are motivated by a major deficiency in Nash equilibrium: it provides virtually no guarantees on how it will play in parts of the game tree that are reached with zero probability. Refinements can mend this issue, but have not been adopted in practice, mostly due to a lack of scalable algorithms. We show that, compared to standard algorithms, our method finds solutions that have substantially better refinement properties, while enjoying a convergence rate that is comparable to that of state-of-the-art algorithms for Nash equilibrium computation both in theory and practice.

****

Coresets for Vector Summarization with Applications to Network Graphs

Dan Feldman, Sedat Ozer, Daniela Rus

We provide a deterministic data summarization algorithm that approximates the mean $\bar{p}=\frac{1}{n}\sum_{p\in P} p$ of a set P of n vectors in $\mathbb{R}^d$, by a weighted mean $\tilde{p}$ of a subset of $O(1/\epsilon)$ vectors, i.e., independent of both n and d. We prove that the squared Euclidean distance between $\bar{p}$ and $\tilde{p}$ is at most ϵ multiplied by the variance of P. We use this algorithm to maintain an approximated sum of vectors from an unbounded stream, using memory that is independent of d, and logarithmic in the n vectors seen so far. Our main application is to extract and represent in a compact way friend groups and activity summaries of users from underlying data exchanges. For example, in the case of mobile networks, we can use GPS traces to identify meetings; in the case of social networks, we can use information exchange to identify friend groups. Our algorithm provably identifies the Heavy Hitter entries in a proximity (adjacency) matrix. The Heavy Hitters can be used to extract and represent in a compact way friend groups and activity summaries of users from underlying data exchanges. We evaluate the algorithm on several large data sets.

****

Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks

Chelsea Finn, Pieter Abbeel, Sergey Levine

We propose an algorithm for meta-learning that is model-agnostic, in the sense that it is compatible with any model trained with gradient descent and applicable to a variety of different learning problems, including classification, regression, and reinforcement learning. The goal of meta-learning is to train a model on a variety of learning tasks, such that it can solve new learning tasks using only a small number of training samples. In our approach, the parameters of the model are explicitly trained such that a small number of gradient steps with a small amount of training data from a new task will produce good generalization performance on that task. In effect, our method trains the model to be easy to fine-tune. We demonstrate that this approach leads to state-of-the-art performance on two few-shot image classification benchmarks, produces good results on few-shot regression, and accelerates fine-tuning for policy gradient reinforcement learning with neural network policies.

****

Input Switched Affine Networks: An RNN Architecture Designed for Interpretability

Jakob N. Foerster, Justin Gilmer, Jascha Sohl-Dickstein, Jan Chorowski, David Sussillo

There exist many problem domains where the interpretability of neural network models is essential for deployment. Here we introduce a recurrent architecture composed of input-switched affine transformations – in other words an RNN without any explicit nonlinearities, but with input-dependent recurrent weights. This simple form allows the RNN to be analyzed via straightforward linear methods: we can exactly characterize the linear contribution of each input to the model predictions; we can use a change-of-basis to disentangle input, output, and computational hidden unit subspaces; we can fully reverse-engineer the architecture’s solution to a simple task. Despite this ease of interpretation, the input switched affine network achieves reasonable performance on a text modeling tasks, and allows greater computational efficiency than networks with standard nonlinearities.

****

Stabilising Experience Replay for Deep Multi-Agent Reinforcement Learning

Jakob Foerster, Nantas Nardelli, Gregory Farquhar, Triantafyllos Afouras, Philip H. S. Torr, Pushmeet Kohli, Shimon Whiteson

Many real-world problems, such as network packet routing and urban traffic control, are naturally modeled as multi-agent reinforcement learning (RL) problems. However, existing multi-agent RL methods typically scale poorly in the problem size. Therefore, a key challenge is to translate the success of deep learning on single-agent RL to the multi-agent setting. A major stumbling block is that independent Q-learning, the most popular multi-agent RL method, introduces nonstationarity that makes it incompatible with the experience replay memory on which deep Q-learning relies. This paper proposes two methods that address this problem: 1) using a multi-agent variant of importance sampling to naturally decay obsolete data and 2) conditioning each agent’s value function on a fingerprint that disambiguates the age of the data sampled from the replay memory. Results on a challenging decentralised variant of StarCraft unit micromanagement confirm that these methods enable the successful combination of experience replay with multi-agent RL.

****

Counterfactual Data-Fusion for Online Reinforcement Learners

Andrew Forney, Judea Pearl, Elias Bareinboim

The Multi-Armed Bandit problem with Unobserved Confounders (MABUC) considers decision-making settings where unmeasured variables can influence both the agent’s decisions and received rewards (Bareinboim et al., 2015). Recent findings showed that unobserved confounders (UCs) pose a unique challenge to algorithms based on standard randomization (i.e., experimental data); if UCs are naively averaged out, these algorithms behave sub-optimally, possibly incurring infinite regret. In this paper, we show how counterfactual-based decision-making circumvents thes e problems and leads to a coherent fusion of observational and experimental data . We then demonstrate this new strategy in an enhanced Thompson Sampling bandit player, and support our findings’ efficacy with extensive simulations.

****

Forward and Reverse Gradient-Based Hyperparameter Optimization

Luca Franceschi, Michele Donini, Paolo Frasconi, Massimiliano Pontil

We study two procedures (reverse-mode and forward-mode) for computing the gradient of the validation error with respect to the hyperparameters of any iterative learning algorithm such as stochastic gradient descent. These procedures mirror two ways of computing gradients for recurrent neural networks and have different trade-offs in terms of running time and space requirements. Our formulation of the reverse-mode procedure is linked to previous work by Maclaurin et al (2015) but does not require reversible dynamics. Additionally, we explore the use of constraints on the hyperparameters. The forward-mode procedure is suitable for real-time hyperparameter updates, which may significantly speedup hyperparameter optimization on large datasets. We present a series of experiments on image and phone classification tasks. In the second task, previous gradient-based approaches are prohibitive. We show that our real-time algorithm yields state-of-the-art results in affordable time.

****

Learning to Detect Sepsis with a Multitask Gaussian Process RNN Classifier

Joseph Futoma, Sanjay Hariharan, Katherine Heller

We present a scalable end-to-end classifier that uses streaming physiological and medication data to accurately predict the onset of sepsis, a life-threatening complication from infections that has high mortality and morbidity. Our proposed framework models the multivariate trajectories of continuous-valued physiological time series using multitask Gaussian processes, seamlessly accounting for the high uncertainty, frequent missingness, and irregular sampling rates typically associated with real clinical data. The Gaussian process is directly connected to a black-box classifier that predicts whether a patient will become septic, chosen in our case to be a recurrent neural network to account for the extreme variability in the length of patient encounters. We show how to scale the computations associated with the Gaussian process in a manner so that the entire system can be discriminatively trained end-to-end using backpropagation. In a large cohort of heterogeneous inpatient encounters at our university health system we find that it outperforms several baselines at predicting sepsis, and yields 19.4\% and 55.5\% improved areas under the Receiver Operating Characteristic and Precision Recall curves as compared to the NEWS score currently used by our hospital.

****

Deep Bayesian Active Learning with Image Data

Yarin Gal, Riashat Islam, Zoubin Ghahramani

Even though active learning forms an important pillar of machine learning, deep learning tools are not prevalent within it. Deep learning poses several difficulties when used in an active learning setting. First, active learning (AL) methods generally rely on being able to learn and update models from small amounts of data. Recent advances in deep learning, on the other hand, are notorious for their dependence on large amounts of data. Second, many AL acquisition functions rely on model uncertainty, yet deep learning methods rarely represent such model uncertainty. In this paper we combine recent advances in Bayesian deep learning into the active learning framework in a practical way. We develop an active learning framework for high dimensional data, a task which has been extremely challenging so far, with very sparse existing literature. Taking advantage of specialised models such as Bayesian convolutional neural networks, we demonstrate our active learning techniques with image data, obtaining a significant improvement on existing active learning approaches. We demonstrate this on both the MNIST dataset, as well as for skin cancer diagnosis from lesion images (ISIC2016 task).

****

Local-to-Global Bayesian Network Structure Learning

Tian Gao, Kshitij Fadnis, Murray Campbell

We introduce a new local-to-global structure learning algorithm, called graph gr owing structure learning (GGSL), to learn Bayesian network (BN) structures. GGSL starts at a (random) node and then gradually expands the learned structure through a series of local learning steps. At each local learning step, the proposed algorithm only needs to revisit a subset of the learned nodes, consisting of the local neighborhood of a target, and therefore improves on both memory and time efficiency compared to traditional global structure learning approaches. GGSL also improves on the existing local-to-global learning approaches by removing the need for conflict-resolving AND-rules, and achieves better learning accuracy. We provide theoretical analysis for the local learning step, and show that GGSL outperforms existing algorithms on benchmark datasets. Overall, GGSL demonstrates a novel direction to scale up BN structure learning while limiting accuracy loss.

****

Communication-efficient Algorithms for Distributed Stochastic Principal Component Analysis

Dan Garber, Ohad Shamir, Nathan Srebro

We study the fundamental problem of Principal Component Analysis in a statistical distributed setting in which each machine out of m stores a sample of n points sampled i.i.d. from a single unknown distribution. We study algorithms for estimating the leading principal component of the population covariance matrix that are both communication-efficient and achieve estimation error of the order of the centralized ERM solution that uses all mn samples. On the negative side, we show that in contrast to results obtained for distributed estimation under convexity assumptions, for the PCA objective, simply averaging the local ERM solutions cannot guarantee error that is consistent with the centralized ERM. We show that this unfortunate phenomena can be remedied by performing a simple correction step which correlates between the individual solutions, and provides an estimator that is consistent with the centralized ERM for sufficiently-large n. We also introduce an iterative distributed algorithm that is applicable in any regime of n, which is based on distributed matrix-vector products. The algorithm gives significant acceleration in terms of communication rounds over previous distributed algorithms, in a wide regime of parameters.

****

Differentiable Programs with Neural Libraries

Alexander L. Gaunt, Marc Brockschmidt, Nate Kushman, Daniel Tarlow

We develop a framework for combining differentiable programming languages with neural networks. Using this framework we create end-to-end trainable systems that learn to write interpretable algorithms with perceptual components. We explore the benefits of inductive biases for strong generalization and modularity that come from the program-like structure of our models. In particular, modularity allows us to learn a library of (neural) functions which grows and improves as more tasks are solved. Empirically, we show that this leads to lifelong learning systems that transfer knowledge to new tasks more effectively than baselines.

****

Zonotope Hit-and-run for Efficient Sampling from Projection DPPs

Guillaume Gautier, Rémi Bardenet, Michal Valko

Determinantal point processes (DPPs) are distributions over sets of items that model diversity using kernels. Their applications in machine learning include summary extraction and recommendation systems. Yet, the cost of sampling from a DPP is prohibitive in large-scale applications, which has triggered an effort towards efficient approximate samplers. We build a novel MCMC sampler that combines ideas from combinatorial geometry, linear programming, and Monte Carlo methods to sample from DPPs with a fixed sample cardinality, also called projection DPPs. Our sampler leverages the ability of the hit-and-run MCMC kernel to efficiently move across convex bodies. Previous theoretical results yield a fast mixing time of our chain when targeting a distribution that is close to a projection DPP, but not a DPP in general. Our empirical results demonstrate that this extends to sampling projection DPPs, i.e., our sampler is more sample-efficient than previous approaches which in turn translates to faster convergence when dealing with costly-to-evaluate functions, such as summary extraction in our experiments.

****

No Spurious Local Minima in Nonconvex Low Rank Problems: A Unified Geometric Analysis

Rong Ge, Chi Jin, Yi Zheng

In this paper we develop a new framework that captures the common landscape underlying the common non-convex low-rank matrix problems including matrix sensing, matrix completion and robust PCA. In particular, we show for all above problems (including asymmetric cases): 1) all local minima are also globally optimal; 2) no high-order saddle points exists. These results explain why simple algorithms such as stochastic gradient descent have global converge, and efficiently optimize these non-convex objective functions in practice. Our framework connects and simplifies the existing analyses on optimization landscapes for matrix sensing and symmetric matrix completion. The framework naturally leads to new results for asymmetric matrix completion and robust PCA.

****

Convolutional Sequence to Sequence Learning

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin

The prevalent approach to sequence to sequence learning maps an input sequence to a variable length output sequence via recurrent neural networks. We introduce an architecture based entirely on convolutional neural networks. Compared to recurrent models, computations over all elements can be fully parallelized during training to better exploit the GPU hardware and optimization is easier since the number of non-linearities is fixed and independent of the input length. Our use of gated linear units eases gradient propagation and we equip each decoder layer with a separate attention module. We outperform the accuracy of the deep LSTM setup of Wu et al. (2016) on both WMT’14 English-German and WMT’14 English-French translation at an order of magnitude faster speed, both on GPU and CPU.

****

On Context-Dependent Clustering of Bandits

Claudio Gentile, Shuai Li, Purushottam Kar, Alexandros Karatzoglou, Giovanni Zappella, Evans Etrue

We investigate a novel cluster-of-bandit algorithm CAB for collaborative recommendation tasks that implements the underlying feedback sharing mechanism by estimating user neighborhoods in a context-dependent manner. CAB makes sharp departures from the state of the art by incorporating collaborative effects into inference, as well as learning processes in a manner that seamlessly interleaves explore-exploit tradeoffs and collaborative steps. We prove regret bounds for CAB under various data-dependent assumptions which exhibit a crisp dependence on the expected number of clusters over the users, a natural measure of the statistical difficulty of the learning task. Experiments on production and real-world datasets show that CAB offers significantly increased prediction performance against a representative pool of state-of-the-art methods.

****

Neural Message Passing for Quantum Chemistry

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, George E. Dahl

Supervised learning on molecules has incredible potential to be useful in chemistry, drug discovery, and materials science. Luckily, several promising and closely related neural network models invariant to molecular symmetries have already been described in the literature. These models learn a message passing algorithm and aggregation procedure to compute a function of their entire input graph. At this point, the next step is to find a particularly effective variant of this general approach and apply it to chemical prediction benchmarks until we either solve them or reach the limits of the approach. In this paper, we reformulate existing models into a single common framework we call Message Passing Neural Networks (MPNNs) and explore additional novel variations within this framework. Using MPNNs we demonstrate state of the art results on an important molecular property prediction benchmark; these results are strong enough that we believe future work should focus on datasets with larger molecules or more accurate ground truth labels.

****

Convex Phase Retrieval without Lifting via PhaseMax

Tom Goldstein, Christoph Studer

Semidefinite relaxation methods transform a variety of non-convex optimization problems into convex problems, but square the number of variables. We study a new type of convex relaxation for phase retrieval problems, called PhaseMax, that convexifies the underlying problem without lifting. The resulting problem formulation can be solved using standard convex optimization routines, while still working in the original, low-dimensional variable space. We prove, using a random spherical distribution measurement model, that PhaseMax succeeds with high probability for a sufficiently large number of measurements. We compare our approach to other phase retrieval methods and demonstrate that our theory accurately predicts the success of PhaseMax.

****

Preferential Bayesian Optimization

Javier González, Zhenwen Dai, Andreas Damianou, Neil D. Lawrence

Bayesian optimization (BO) has emerged during the last few years as an effective approach to optimize black-box functions where direct queries of the objective are expensive. We consider the case where direct access to the function is not possible, but information about user preferences is. Such scenarios arise in problems where human preferences are modeled, such as A/B tests or recommender systems. We present a new framework for this scenario that we call Preferential Bayesian Optimization (PBO) and that allows to find the optimum of a latent function that can only be queried through pairwise comparisons, so-called duels. PBO extend the applicability of standard BO ideas and generalizes previous discrete dueling approaches by modeling the probability of the the winner of each duel by means of Gaussian process model with a Bernoulli likelihood. The latent preference function is used to define a family of acquisition functions that extend usual policies used in BO. We illustrate the benefits of PBO in a variety of experiments in which we show how the way correlations are modeled is the key ingredient to drastically reduce the number of comparisons to find the optimum of the latent function of interest.

****

Measuring Sample Quality with Kernels

Jackson Gorham, Lester Mackey

Approximate Markov chain Monte Carlo (MCMC) offers the promise of more rapid sampling at the cost of more biased inference. Since standard MCMC diagnostics fail to detect these biases, researchers have developed computable Stein discrepancy measures that provably determine the convergence of a sample to its target distribution. This approach was recently combined with the theory of reproducing kernels to define a closed-form kernel Stein discrepancy (KSD) computable by summing kernel evaluations across pairs of sample points. We develop a theory of weak convergence for KSDs based on Stein’s method, demonstrate that commonly used KSDs fail to detect non-convergence even for Gaussian targets, and show that kernels with slowly decaying tails provably determine convergence for a large class of target distributions. The resulting convergence-determining KSDs are suitable for comparing biased, exact, and deterministic sample sequences and simpler to compute and parallelize than alternative Stein discrepancies. We use our tools to compare biased samplers, select sampler hyperparameters, and improve upon existing KSD approaches to one-sample hypothesis testing and sample quality improvement.

****

Efficient softmax approximation for GPUs

Grave, Armand Joulin, Moustapha Cissé, David Grangier, Hervé Jégou

We propose an approximate strategy to efficiently train neural network based language models over very large vocabularies. Our approach, called adaptive softmax, circumvents the linear dependency on the vocabulary size by exploiting the unbalanced word distribution to form clusters that explicitly minimize the expectation of computation time. Our approach further reduces the computational cost by exploiting the specificities of modern architectures and matrix-matrix vector op erations, making it particularly suited for graphical processing units. Our experiments carried out on standard benchmarks, such as EuroParl and One Billion Word, show that our approach brings a large gain in efficiency over standard approximations while achieving an accuracy close to that of the full softmax. The code of our method is available at https://github.com/facebookresearch/adaptive-softmax.

****

Automated Curriculum Learning for Neural Networks

Alex Graves, Marc G. Bellemare, Jacob Menick, Rémi Munos, Koray Kavukcuoglu

We introduce a method for automatically selecting the path, or syllabus, that a neural network follows through a curriculum so as to maximise learning efficiency. A measure of the amount that the network learns from each data sample is provided as a reward signal to a nonstationary multi-armed bandit algorithm, which then determines a stochastic syllabus. We consider a range of signals derived from two distinct indicators of learning progress: rate of increase in prediction accuracy, and rate of increase in network complexity. Experimental results for LSTM networks on three curricula demonstrate that our approach can significantly accelerate learning, in some cases halving the time required to attain a satisfactory performance level.

****

On Calibration of Modern Neural Networks

Chuan Guo, Geoff Pleiss, Yu Sun, Kilian Q. Weinberger

Confidence calibration – the problem of predicting probability estimates representative of the true correctness likelihood – is important for classification models in many applications. We discover that modern neural networks, unlike those from a decade ago, are poorly calibrated. Through extensive experiments, we observe that depth, width, weight decay, and Batch Normalization are important factors influencing calibration. We evaluate the performance of various post-processing calibration methods on state-of-the-art architectures with image and document classification datasets. Our analysis and experiments not only offer insights into neural network learning, but also provide a simple and straightforward recipe for practical settings: on most datasets, temperature scaling – a single-parameter variant of Platt Scaling – is surprisingly effective at calibrating predictions.

****

ProtoNN: Compressed and Accurate kNN for Resource-scarce Devices

Chirag Gupta, Arun Sai Suggala, Ankit Goyal, Harsha Vardhan Simhadri, Bhargavi Paranjape, Ashish Kumar, Saurabh Goyal, Raghavendra Udupa, Manik Varma, Prateek Jain

Several real-world applications require real-time prediction on resource-scarce devices such as an Internet of Things (IoT) sensor. Such applications demand prediction models with small storage and computational complexity that do not compromise significantly on accuracy. In this work, we propose ProtoNN, a novel algorithm that addresses the problem of real-time and accurate prediction on resource-scarce devices. ProtoNN is inspired by k-Nearest Neighbor (KNN) but has several orders lower storage and prediction complexity. ProtoNN models can be deployed even on devices with puny storage and computational power (e.g. an Arduino UNO with 2kB RAM) to get excellent prediction accuracy. ProtoNN derives its strength from three key ideas: a) learning a small number of prototypes to represent the entire training set, b) sparse low dimensional projection of data, c) joint discriminative learning of the projection and prototypes with explicit model size constraint. We conduct systematic empirical evaluation of ProtoNN on a variety of supervised learning tasks (binary, multi-class, multi-label classification) and show that it gives nearly state-of-the-art prediction accuracy on resource-scarce devices while consuming several orders lower storage, and using minimal working memory.

****

Deep Value Networks Learn to Evaluate and Iteratively Refine Structured Outputs

Michael Gygli, Mohammad Norouzi, Anelia Angelova

We approach structured output prediction by optimizing a deep value network (DVN

) to precisely estimate the task loss on different output configurations for a given input. Once the model is trained, we perform inference by gradient descent on the continuous relaxations of the output variables to find outputs with promising scores from the value network. When applied to image segmentation, the value network takes an image and a segmentation mask as inputs and predicts a scalar estimating the intersection over union between the input and ground truth masks. For multi-label classification, the DVN’s objective is to correctly predict the F1 score for any potential label configuration. The DVN framework achieves the state-of-the-art results on multi-label prediction and image segmentation benchmarks.

****

Reinforcement Learning with Deep Energy-Based Policies

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, Sergey Levine

We propose a method for learning expressive energy-based policies for continuous states and actions, which has been feasible only in tabular domains before. We apply our method to learning maximum entropy policies, resulting into a new algorithm, called soft Q-learning, that expresses the optimal policy via a Boltzmann distribution. We use the recently proposed amortized Stein variational gradient descent to learn a stochastic sampling network that approximates samples from this distribution. The benefits of the proposed algorithm include improved exploration and compositionality that allows transferring skills between tasks, which we confirm in simulated experiments with swimming and walking robots. We also draw a connection to actor-critic methods, which can be viewed performing approximate inference on the corresponding energy-based model.

****

DeepBach: a Steerable Model for Bach Chorales Generation

Gaëtan Hadjeres, François Pachet, Frank Nielsen

This paper introduces DeepBach, a graphical model aimed at modeling polyphonic music and specifically hymn-like pieces. We claim that, after being trained on the chorale harmonizations by Johann Sebastian Bach, our model is capable of generating highly convincing chorales in the style of Bach. DeepBach’s strength comes from the use of pseudo-Gibbs sampling coupled with an adapted representation of musical data. This is in contrast with many automatic music composition approaches which tend to compose music sequentially. Our model is also steerable in the sense that a user can constrain the generation by imposing positional constraints such as notes, rhythms or cadences in the generated score. We also provide a plugin on top of the MuseScore music editor making the interaction with DeepBach easy to use.

****

Consistent On-Line Off-Policy Evaluation

Assaf Hallak, Shie Mannor

The problem of on-line off-policy evaluation (OPE) has been actively studied in the last decade due to its importance both as a stand-alone problem and as a module in a policy improvement scheme. However, most Temporal Difference (TD) based solutions ignore the discrepancy between the stationary distribution of the behavior and target policies and its effect on the convergence limit when function approximation is applied. In this paper we propose the Consistent Off-Policy Temporal Difference (COP-TD(λ, β)) algorithm that addresses this issue and reduces this bias at some computational expense. We show that COP-TD(λ, β) can be designed to converge to the same value that would have been obtained by using on-policy TD(λ) with the target policy. Subsequently, the proposed scheme leads to a related and promising heuristic we call log-COP-TD(λ, β). Both algorithms have favorable empirical results to the current state of the art on-line OPE algorithms. Finally, our formulation sheds some new light on the recently proposed Emphatic TD learning.

****

Faster Greedy MAP Inference for Determinantal Point Processes

Insu Han, Prabhanjan Kambadur, Kyoungsoo Park, Jinwoo Shin

Determinantal point processes (DPPs) are popular probabilistic models that arise in many machine learning tasks, where distributions of diverse sets are charact erized by determinants of their features. In this paper, we develop fast algorithms to find the most likely configuration (MAP) of large-scale DPPs, which is NP-hard in general. Due to the submodular nature of the MAP objective, greedy algorithms have been used with empirical success. Greedy implementations require computation of log-determinants, matrix inverses or solving linear systems at each iteration. We present faster implementations of the greedy algorithms by utilizing the orthogonal benefits of two log-determinant approximation schemes: (a) first-order expansions to the matrix log-determinant function and (b) high-order expansions to the scalar log function with stochastic trace estimators. In our experiments, our algorithms are orders of magnitude faster than their competitors, while sacrificing marginal accuracy.

****

Data-Efficient Policy Evaluation Through Behavior Policy Search

Josiah P. Hanna, Philip S. Thomas, Peter Stone, Scott Niekum

We consider the task of evaluating a policy for a Markov decision process (MDP). The standard unbiased technique for evaluating a policy is to deploy the policy and observe its performance. We show that the data collected from deploying a different policy, commonly called the behavior policy, can be used to produce unbiased estimates with lower mean squared error than this standard technique. We derive an analytic expression for the optimal behavior policy — the behavior policy that minimizes the mean squared error of the resulting estimates. Because this expression depends on terms that are unknown in practice, we propose a novel policy evaluation sub-problem, behavior policy search: searching for a behavior policy that reduces mean squared error. We present a behavior policy search algorithm and empirically demonstrate its effectiveness in lowering the mean squared error of policy performance estimates.

****

Joint Dimensionality Reduction and Metric Learning: A Geometric Take

Mehrtash Harandi, Mathieu Salzmann, Richard Hartley

To be tractable and robust to data noise, existing metric learning algorithms commonly rely on PCA as a pre-processing step. How can we know, however, that PCA, or any other specific dimensionality reduction technique, is the method of choice for the problem at hand? The answer is simple: We cannot! To address this issue, in this paper, we develop a Riemannian framework to jointly learn a mapping performing dimensionality reduction and a metric in the induced space. Our experiments evidence that, while we directly work on high-dimensional features, our approach yields competitive runtimes with and higher accuracy than state-of-the-art metric learning algorithms.

****

Deep IV: A Flexible Approach for Counterfactual Prediction

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, Matt Taddy

Counterfactual prediction requires understanding causal relationships between so-called treatment and outcome variables. This paper provides a recipe for augmenting deep learning methods to accurately characterize such relationships in the presence of instrument variables (IVs) – sources of treatment randomization that are conditionally independent from the outcomes. Our IV specification resolves into two prediction tasks that can be solved with deep neural nets: a first-stage network for treatment prediction and a second-stage network whose loss function involves integration over the conditional treatment distribution. This Deep IV framework allows us to take advantage of off-the-shelf supervised learning techniques to estimate causal effects by adapting the loss function. Experiments show that it outperforms existing machine learning approaches.

****

Robust Guarantees of Stochastic Greedy Algorithms

Avinatan Hassidim, Yaron Singer

In this paper we analyze the robustness of stochastic variants of the greedy algorithm for submodular maximization. Our main result shows that for maximizing a monotone submodular function under a cardinality constraint, iteratively selecting an element whose marginal contribution is approximately maximal in expectation is a sufficient condition to obtain the optimal approximation guarantee with e

xponentially high probability, assuming the cardinality is sufficiently large. One consequence of our result is that the linear-time STOCHASTIC-GREEDY algorithm recently proposed in (Mirzasoleiman et al.,2015) achieves the optimal running time while maintaining an optimal approximation guarantee. We also show that high probability guarantees cannot be obtained for stochastic greedy algorithms under matroid constraints, and prove an approximation guarantee which holds in expectation. In contrast to the guarantees of the greedy algorithm, we show that the approximation ratio of stochastic local search is arbitrarily bad, with high probability, as well as in expectation.
****
Efficient Regret Minimization in Non-Convex Games
Elad Hazan, Karan Singh, Cyril Zhang
We consider regret minimization in repeated games with non-convex loss functions. Minimizing the standard notion of regret is computationally intractable. Thus, we define a natural notion of regret which permits efficient optimization and generalizes offline guarantees for convergence to an approximate local optimum. We give gradient-based methods that achieve optimal regret, which in turn guarantee convergence to equilibrium in this framework.
****
Kernelized Support Tensor Machines
Lifang He, Chun-Ta Lu, Guixiang Ma, Shen Wang, Linlin Shen, Philip S. Yu, Ann B. Ragin
In the context of supervised tensor learning, preserving the structural information and exploiting the discriminative nonlinear relationships of tensor data are crucial for improving the performance of learning tasks. Based on tensor factorization theory and kernel methods, we propose a novel Kernelized Support Tensor Machine (KSTM) which integrates kernelized tensor factorization with maximum-margin criterion. Specifically, the kernelized factorization technique is introduced to approximate the tensor data in kernel space such that the complex nonlinear relationships within tensor data can be explored. Further, dual structural preserving kernels are devised to learn the nonlinear boundary between tensor data. As a result of joint optimization, the kernels obtained in KSTM exhibit better generalization power to discriminative analysis. The experimental results on real-world neuroimaging datasets show the superiority of KSTM over the state-of-the-art techniques.
****
The Sample Complexity of Online One-Class Collaborative Filtering
Reinhard Heckel, Kannan Ramchandran
We consider the online one-class collaborative filtering (CF) problem that consist of recommending items to users over time in an online fashion based on positive ratings only. This problem arises when users respond only occasionally to a recommendation with a positive rating, and never with a negative one. We study the impact of the probability of a user responding to a recommendation, p_f, on the sample complexity, and ask whether receiving positive and negative ratings, instead of positive ratings only, improves the sample complexity. Both questions arise in the design of recommender systems. We introduce a simple probabilistic user model, and analyze the performance of an online user-based CF algorithm. We prove that after an initial cold start phase, where recommendations are invested in exploring the user’s preferences, this algorithm makes—up to a fraction of the recommendations required for updating the user’s preferences—perfect recommendations. The number of ratings required for the cold start phase is nearly proportional to $1/p_f$, and that for updating the user’s preferences is essentially independent of p_f. As a consequence we find that, receiving positive and negative ratings instead of only positive ones improves the number of ratings required for initial exploration by a factor of $1/p_f$, which can be significant.
****
Warped Convolutions: Efficient Invariance to Spatial Transformations
João F. Henriques, Andrea Vedaldi
Convolutional Neural Networks (CNNs) are extremely efficient, since they exploit the inherent translation-invariance of natural images. However, translation is

just one of a myriad of useful spatial transformations. Can the same efficiency be attained when considering other spatial invariances? Such generalized convolutions have been considered in the past, but at a high computational cost. We present a construction that is simple and exact, yet has the same computational complexity that standard convolutions enjoy. It consists of a constant image warp followed by a simple convolution, which are standard blocks in deep learning toolboxes. With a carefully crafted warp, the resulting architecture can be made equivariant to a wide range of two-parameter spatial transformations. We show encouraging results in realistic scenarios, including the estimation of vehicle poses in the Google Earth dataset (rotation and scale), and face poses in Annotated Facial Landmarks in the Wild (3D rotations under perspective).

****

Parallel and Distributed Thompson Sampling for Large-scale Accelerated Exploration of Chemical Space

José Miguel Hernández-Lobato, James Requeima, Edward O. Pyzer-Knapp, Alán Aspuru-Guzik

Chemical space is so large that brute force searches for new interesting molecules are infeasible. High-throughput virtual screening via computer cluster simulations can speed up the discovery process by collecting very large amounts of data in parallel, e.g., up to hundreds or thousands of parallel measurements. Bayesian optimization (BO) can produce additional acceleration by sequentially identifying the most useful simulations or experiments to be performed next. However, current BO methods cannot scale to the large numbers of parallel measurements and the massive libraries of molecules currently used in high-throughput screening. Here, we propose a scalable solution based on a parallel and distributed implementation of Thompson sampling (PDTS). We show that, in small scale problems, PDTS performs similarly as parallel expected improvement (EI), a batch version of the most widely used BO heuristic. Additionally, in settings where parallel EI does not scale, PDTS outperforms other scalable baselines such as a greedy search, ϵ-greedy approaches and a random search method. These results show that PDTS is a successful solution for large-scale parallel BO.

****

DARLA: Improving Zero-Shot Transfer in Reinforcement Learning

Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, Alexander Lerchner

Domain adaptation is an important open problem in deep reinforcement learning (RL). In many scenarios of interest data is hard to obtain, so agents may learn a source policy in a setting where data is readily available, with the hope that it generalises well to the target domain. We propose a new multi-stage RL agent, DARLA (DisentAngled Representation Learning Agent), which learns to see before learning to act. DARLA’s vision is based on learning a disentangled representation of the observed environment. Once DARLA can see, it is able to acquire source policies that are robust to many domain shifts – even with no access to the target domain. DARLA significantly outperforms conventional baselines in zero-shot domain adaptation scenarios, an effect that holds across a variety of RL environments (Jaco arm, DeepMind Lab) and base RL algorithms (DQN, A3C and EC).

****

SPLICE: Fully Tractable Hierarchical Extension of ICA with Pooling

Jun-ichiro Hirayama, Aapo Hyvärinen, Motoaki Kawanabe

We present a novel probabilistic framework for a hierarchical extension of independent component analysis (ICA), with a particular motivation in neuroscientific data analysis and modeling. The framework incorporates a general subspace pooling with linear ICA-like layers stacked recursively. Unlike related previous models, our generative model is fully tractable: both the likelihood and the posterior estimates of latent variables can readily be computed with analytically simple formulae. The model is particularly simple in the case of complex-valued data since the pooling can be reduced to taking the modulus of complex numbers. Experiments on electroencephalography (EEG) and natural images demonstrate the validity of the method.

****

Multilevel Clustering via Wasserstein Means

Nhat Ho, XuanLong Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, Dinh Phung

We propose a novel approach to the problem of multilevel clustering, which aims to simultaneously partition data in each group and discover grouping patterns among groups in a potentially large hierarchically structured corpus of data. Our method involves a joint optimization formulation over several spaces of discrete probability measures, which are endowed with Wasserstein distance metrics. We propose a number of variants of this problem, which admit fast optimization algorithms, by exploiting the connection to the problem of finding Wasserstein barycenters. Consistency properties are established for the estimates of both local and global clusters. Finally, experiment results with both synthetic and real data are presented to demonstrate the flexibility and scalability of the proposed approach.

****

Learning Deep Latent Gaussian Models with Markov Chain Monte Carlo

Matthew D. Hoffman

Deep latent Gaussian models are powerful and popular probabilistic models of high-dimensional data. These models are almost always fit using variational expectation-maximization, an approximation to true maximum-marginal-likelihood estimation. In this paper, we propose a different approach: rather than use a variational approximation (which produces biased gradient signals), we use Markov chain Monte Carlo (MCMC, which allows us to trade bias for computation). We find that our MCMC-based approach has several advantages: it yields higher held-out likelihoods, produces sharper images, and does not suffer from the variational overpruning effect. MCMC’s additional computational overhead proves to be significant, but not prohibitive.

****

Minimizing Trust Leaks for Robust Sybil Detection

János Höner, Shinichi Nakajima, Alexander Bauer, Klaus-Robert Müller, Nico Görnitz

Sybil detection is a crucial task to protect online social networks (OSNs) against intruders who try to manipulate automatic services provided by OSNs to their customers. In this paper, we first discuss the robustness of graph-based Sybil detectors SybilRank and Integro and refine theoretically their security guarantees towards more realistic assumptions. After that, we formally introduce adversarial settings for the graph-based Sybil detection problem and derive a corresponding optimal attacking strategy by exploitation of trust leaks. Based on our analysis, we propose transductive Sybil ranking (TSR), a robust extension to SybilRank and Integro that directly minimizes trust leaks. Our empirical evaluation shows significant advantages of TSR over state-of-the-art competitors on a variety of attacking scenarios on artificially generated data and real-world datasets.

****

Prox-PDA: The Proximal Primal-Dual Algorithm for Fast Distributed Nonconvex Optimization and Learning Over Networks

Mingyi Hong, Davood Hajinezhad, Ming-Min Zhao

In this paper we consider nonconvex optimization and learning over a network of distributed nodes. We develop a Proximal Primal-Dual Algorithm (Prox-PDA), which enables the network nodes to distributedly and collectively compute the set of first-order stationary solutions in a global sublinear manner [with a rate of $O(1/r)$, where r is the iteration counter]. To the best of our knowledge, this is the first algorithm that enables distributed nonconvex optimization with global rate guarantees. Our numerical experiments also demonstrate the effectiveness of the proposed algorithm.

****

Analysis and Optimization of Graph Decompositions by Lifted Multicuts

Andrea Hor■áková, Jan-Hendrik Lange, Bjoern Andres

We study the set of all decompositions (clusterings) of a graph through its characterization as a set of lifted multicuts. This leads us to practically relevant insights related to the definition of classes of decompositions by must-join an d must-cut constraints and related to the comparison of clusterings by metrics. To find optimal decompositions defined by minimum cost lifted multicuts, we establish some properties of some facets of lifted multicut polytopes, define efficient separation procedures and apply these in a branch-and-cut algorithm.

****

Dissipativity Theory for Nesterov’s Accelerated Method

Bin Hu, Laurent Lessard

In this paper, we adapt the control theoretic concept of dissipativity theory to provide a natural understanding of Nesterov’s accelerated method. Our theory ties rigorous convergence rate analysis to the physically intuitive notion of energy dissipation. Moreover, dissipativity allows one to efficiently construct Lyapunov functions (either numerically or analytically) by solving a small semidefinite program. Using novel supply rate functions, we show how to recover known rate bounds for Nesterov’s method and we generalize the approach to certify both linear and sublinear rates in a variety of settings. Finally, we link the continuous-time version of dissipativity to recent works on algorithm analysis that use discretizations of ordinary differential equations.

****

Learning Discrete Representations via Information Maximizing Self-Augmented Training

Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, Masashi Sugiyama

Learning discrete representations of data is a central machine learning task because of the compactness of the representations and ease of interpretation. The task includes clustering and hash learning as special cases. Deep neural networks are promising to be used because they can model the non-linearity of data and scale to large datasets. However, their model complexity is huge, and therefore, we need to carefully regularize the networks in order to learn useful representations that exhibit intended invariance for applications of interest. To this end, we propose a method called Information Maximizing Self-Augmented Training (IMSAT). In IMSAT, we use data augmentation to impose the invariance on discrete representations. More specifically, we encourage the predicted representations of augmented data points to be close to those of the original data points in an end-to-end fashion. At the same time, we maximize the information-theoretic dependency between data and their predicted discrete representations. Extensive experiments on benchmark datasets show that IMSAT produces state-of-the-art results for both clustering and unsupervised hash learning.

****

State-Frequency Memory Recurrent Neural Networks

Hao Hu, Guo-Jun Qi

Modeling temporal sequences plays a fundamental role in various modern applications and has drawn more and more attentions in the machine learning community. Among those efforts on improving the capability to represent temporal data, the Long Short-Term Memory (LSTM) has achieved great success in many areas. Although the LSTM can capture long-range dependency in the time domain, it does not explicitly model the pattern occurrences in the frequency domain that plays an important role in tracking and predicting data points over various time cycles. We propose the State-Frequency Memory (SFM), a novel recurrent architecture that allows to separate dynamic patterns across different frequency components and their impacts on modeling the temporal contexts of input sequences. By jointly decomposing memorized dynamics into state-frequency components, the SFM is able to offer a fine-grained analysis of temporal sequences by capturing the dependency of uncovered patterns in both time and frequency domains. Evaluations on several temporal modeling tasks demonstrate the SFM can yield competitive performances, in particular as compared with the state-of-the-art LSTM models.

****

Deep Generative Models for Relational Data with Side Information

Changwei Hu, Piyush Rai, Lawrence Carin

We present a probabilistic framework for overlapping community discovery and link prediction for relational data, given as a graph. The proposed framework has: (1) a deep architecture which enables us to infer multiple layers of latent feat ures/communities for each node, providing superior link prediction performance on more complex networks and better interpretability of the latent features; and (2) a regression model which allows directly conditioning the node latent features on the side information available in form of node attributes. Our framework handles both (1) and (2) via a clean, unified model, which enjoys full local conjugacy via data augmentation, and facilitates efficient inference via closed form Gibbs sampling. Moreover, inference cost scales in the number of edges which is attractive for massive but sparse networks. Our framework is also easily extendable to model weighted networks with count-valued edges. We compare with various state-of-the-art methods and report results, both quantitative and qualitative, on several benchmark data sets.

****

Toward Controlled Generation of Text

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, Eric P. Xing

Generic generation and manipulation of text is challenging and has limited success compared to recent deep generative modeling in visual domain. This paper aims at generating plausible text sentences, whose attributes are controlled by learning disentangled latent representations with designated semantics. We propose a new neural generative model which combines variational auto-encoders (VAEs) and holistic attribute discriminators for effective imposition of semantic structures. The model can alternatively be seen as enhancing VAEs with the wake-sleep algorithm for leveraging fake samples as extra training data. With differentiable approximation to discrete text samples, explicit constraints on independent attribute controls, and efficient collaborative learning of generator and discriminators, our model learns interpretable representations from even only word annotations, and produces short sentences with desired attributes of sentiment and tenses. Quantitative experiments using trained classifiers as evaluators validate the accuracy of sentence and attribute generation.

****

Tensor Decomposition with Smoothness

Masaaki Imaizumi, Kohei Hayashi

Real data tensors are usually high dimensional but their intrinsic information is preserved in low-dimensional space, which motivates to use tensor decompositions such as Tucker decomposition. Often, real data tensors are not only low dimensional, but also smooth, meaning that the adjacent elements are similar or continuously changing, which typically appear as spatial or temporal data. To incorporate the smoothness property, we propose the smoothed Tucker decomposition (STD). STD leverages the smoothness by the sum of a few basis functions, which reduces the number of parameters. The objective function is formulated as a convex problem and, to solve that, an algorithm based on the alternating direction method of multipliers is derived. We theoretically show that, under the smoothness assumption, STD achieves a better error bound. The theoretical result and performances of STD are numerically verified.

****

Variational Inference for Sparse and Undirected Models

John Ingraham, Debora Marks

Undirected graphical models are applied in genomics, protein structure prediction, and neuroscience to identify sparse interactions that underlie discrete data. Although Bayesian methods for inference would be favorable in these contexts, they are rarely used because they require doubly intractable Monte Carlo sampling. Here, we develop a framework for scalable Bayesian inference of discrete undirected models based on two new methods. The first is Persistent VI, an algorithm for variational inference of discrete undirected models that avoids doubly intractable MCMC and approximations of the partition function. The second is Fadeout, a reparameterization approach for variational inference under sparsity-inducing priors that captures a posteriori correlations between parameters and hyperparameters with noncentered parameterizations. We find that, together, these methods for variational inference substantially improve learning of sparse undirected graphical models in simulated and real problems from physics and biology.

****

Fairness in Reinforcement Learning
Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Aaron Roth
We initiate the study of fairness in reinforcement learning, where the actions of a learning algorithm may affect its environment and future rewards. Our fairness constraint requires that an algorithm never prefers one action over another if the long-term (discounted) reward of choosing the latter action is higher. Our first result is negative: despite the fact that fairness is consistent with the optimal policy, any learning algorithm satisfying fairness must take time exponential in the number of states to achieve non-trivial approximation to the optimal policy. We then provide a provably fair polynomial time algorithm under an approximate notion of fairness, thus establishing an exponential gap between exact and approximate fairness.

****

Decoupled Neural Interfaces using Synthetic Gradients
Max Jaderberg, Wojciech Marian Czarnecki, Simon Osindero, Oriol Vinyals, Alex Graves, David Silver, Koray Kavukcuoglu
Training directed neural networks typically requires forward-propagating data through a computation graph, followed by backpropagating error signal, to produce weight updates. All layers, or more generally, modules, of the network are therefore locked, in the sense that they must wait for the remainder of the network to execute forwards and propagate error backwards before they can be updated. In this work we break this constraint by decoupling modules by introducing a model of the future computation of the network graph. These models predict what the result of the modelled subgraph will produce using only local information. In particular we focus on modelling error gradients: by using the modelled synthetic gradient in place of true backpropagated error gradients we decouple subgraphs, and can update them independently and asynchronously i.e. we realise decoupled neural interfaces. We show results for feed-forward models, where every layer is trained asynchronously, recurrent neural networks (RNNs) where predicting one’s future gradient extends the time over which the RNN can effectively model, and also a hierarchical RNN system with ticking at different timescales. Finally, we demonstrate that in addition to predicting gradients, the same framework can be used to predict inputs, resulting in models which are decoupled in both the forward and backwards pass – amounting to independent networks which co-learn such that they can be composed into a single functioning corporation.

****

Scalable Generative Models for Multi-label Learning with Missing Labels
Vikas Jain, Nirbhay Modhe, Piyush Rai
We present a scalable, generative framework for multi-label learning with missing labels. Our framework consists of a latent factor model for the binary label matrix, which is coupled with an exposure model to account for label missingness (i.e., whether a zero in the label matrix is indeed a zero or denotes a missing observation). The underlying latent factor model also assumes that the low-dimensional embeddings of each label vector are directly conditioned on the respective feature vector of that example. Our generative framework admits a simple inference procedure, such that the parameter estimation reduces to a sequence of simple weighted least-square regression problems, each of which can be solved easily, efficiently, and in parallel. Moreover, inference can also be performed in an online fashion using mini-batches of training examples, which makes our framework scalable for large data sets, even when using moderate computational resources. We report both quantitative and qualitative results for our framework on several benchmark data sets, comparing it with a number of state-of-the-art methods.

****

Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-control
Natasha Jaques, Shixiang Gu, Dzmitry Bahdanau, José Miguel Hernández-Lobato, Richard E. Turner, Douglas Eck
This paper proposes a general method for improving the structure and quality of sequences generated by a recurrent neural network (RNN), while maintaining information originally learned from data, as well as sample diversity. An RNN is firs

t pre-trained on data using maximum likelihood estimation (MLE), and the probability distribution over the next token in the sequence learned by this model is treated as a prior policy. Another RNN is then trained using reinforcement learning (RL) to generate higher-quality outputs that account for domain-specific incentives while retaining proximity to the prior policy of the MLE RNN. To formalize this objective, we derive novel off-policy RL methods for RNNs from KL-control. The effectiveness of the approach is demonstrated on two applications; 1) generating novel musical melodies, and 2) computational molecular generation. For both problems, we show that the proposed method improves the desired properties and structure of the generated sequences, while maintaining information learned from data.

****

Bayesian Optimization with Tree-structured Dependencies

Rodolphe Jenatton, Cedric Archambeau, Javier González, Matthias Seeger

Bayesian optimization has been successfully used to optimize complex black-box functions whose evaluations are expensive. In many applications, like in deep learning and predictive analytics, the optimization domain is itself complex and structured. In this work, we focus on use cases where this domain exhibits a known dependency structure. The benefit of leveraging this structure is twofold: we explore the search space more efficiently and posterior inference scales more favorably with the number of observations than Gaussian Process-based approaches published in the literature. We introduce a novel surrogate model for Bayesian optimization which combines independent Gaussian Processes with a linear model that encodes a tree-based dependency structure and can transfer information between overlapping decision sequences. We also design a specialized two-step acquisition function that explores the search space more effectively. Our experiments on synthetic tree-structured functions and the tuning of feedforward neural networks trained on a range of binary classification datasets show that our method compares favorably with competing approaches.

****

Simultaneous Learning of Trees and Representations for Extreme Classification and Density Estimation

Yacine Jernite, Anna Choromanska, David Sontag

We consider multi-class classification where the predictor has a hierarchical structure that allows for a very large number of labels both at train and test time. The predictive power of such models can heavily depend on the structure of the tree, and although past work showed how to learn the tree structure, it expected that the feature vectors remained static. We provide a novel algorithm to simultaneously perform representation learning for the input data and learning of the hierarchical predictor. Our approach optimizes an objective function which favors balanced and easily-separable multi-way node partitions. We theoretically analyze this objective, showing that it gives rise to a boosting style property and a bound on classification error. We next show how to extend the algorithm to conditional density estimation. We empirically validate both variants of the algorithm on text classification and language modeling, respectively, and show that they compare favorably to common baselines in terms of accuracy and running time.

****

From Patches to Images: A Nonparametric Generative Model

Geng Ji, Michael C. Hughes, Erik B. Sudderth

We propose a hierarchical generative model that captures the self-similar structure of image regions as well as how this structure is shared across image collections. Our model is based on a novel, variational interpretation of the popular expected patch log-likelihood (EPLL) method as a model for randomly positioned grids of image patches. While previous EPLL methods modeled image patches with finite Gaussian mixtures, we use nonparametric Dirichlet process (DP) mixtures to create models whose complexity grows as additional images are observed. An extension based on the hierarchical DP then captures repetitive and self-similar structure via image-specific variations in cluster frequencies. We derive a structured variational inference algorithm that adaptively creates new patch clusters to more accurately model novel image textures. Our denoising performance on standard benchmarks is superior to EPLL and comparable to the state-of-the-art, and provides novel statistical justifications for common image processing heuristics. We also show accurate image inpainting results.

****

Density Level Set Estimation on Manifolds with DBSCAN

Heinrich Jiang

We show that DBSCAN can estimate the connected components of the λ-density level set $\{ x : f(x) \ge \lambda\}$ given n i.i.d. samples from an unknown density f. We characterize the regularity of the level set boundaries using parameter $\beta > 0$ and analyze the estimation error under the Hausdorff metric. When the data lies in $\mathbb{R}^D$ we obtain a rate of $\widetilde{O}(n^{-1/(2\beta + D)})$, which matches known lower bounds up to logarithmic factors. When the data lies on an embedded unknown d-dimensional manifold in $\mathbb{R}^D$, then we obtain a rate of $\widetilde{O}(n^{-1/(2\beta + d\cdot \max\{1, \beta \})})$. Finally, we provide adaptive parameter tuning in order to attain these rates with no a priori knowledge of the intrinsic dimension, density, or β.

****

Uniform Convergence Rates for Kernel Density Estimation

Heinrich Jiang

Kernel density estimation (KDE) is a popular nonparametric density estimation method. We (1) derive finite-sample high-probability density estimation bounds for multivariate KDE under mild density assumptions which hold uniformly in $x \in \mathbb{R}^d$ and bandwidth matrices. We apply these results to (2) mode, (3) density level set, and (4) class probability estimation and attain optimal rates up to logarithmic factors. We then (5) provide an extension of our results under the manifold hypothesis. Finally, we (6) give uniform convergence results for local intrinsic dimension estimation.

****

Contextual Decision Processes with low Bellman rank are PAC-Learnable

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, Robert E. Schapire

This paper studies systematic exploration for reinforcement learning (RL) with rich observations and function approximation. We introduce contextual decision processes (CDPs), that unify most prior RL settings. Our first contribution is a complexity measure, the Bellman rank, that we show enables tractable learning of near-optimal behavior in CDPs and is naturally small for many well-studied RL models. Our second contribution is a new RL algorithm that does systematic exploration to learn near-optimal behavior in CDPs with low Bellman rank. The algorithm requires a number of samples that is polynomial in all relevant parameters but independent of the number of unique contexts. Our approach uses Bellman error minimization with optimistic exploration and provides new insights into efficient exploration for RL with function approximation.

****

Efficient Nonmyopic Active Search

Shali Jiang, Gustavo Malkomes, Geoff Converse, Alyssa Shofner, Benjamin Moseley, Roman Garnett

Active search is an active learning setting with the goal of identifying as many members of a given class as possible under a labeling budget. In this work, we first establish a theoretical hardness of active search, proving that no polynomial-time policy can achieve a constant factor approximation ratio with respect to the expected utility of the optimal policy. We also propose a novel, computationally efficient active search policy achieving exceptional performance on several real-world tasks. Our policy is nonmyopic, always considering the entire remaining search budget. It also automatically and dynamically balances exploration and exploitation consistent with the remaining budget, without relying on a parameter to control this tradeoff. We conduct experiments on diverse datasets from several domains: drug discovery, materials science, and a citation network. Our efficient nonmyopic policy recovers significantly more valuable points with the same budget than several alternatives from the literature, including myopic approximations to the optimal policy.

****

How to Escape Saddle Points Efficiently

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, Michael I. Jordan

This paper shows that a perturbed form of gradient descent converges to a second-order stationary point in a number iterations which depends only poly-logarithmically on dimension (i.e., it is almost “dimension-free”). The convergence rate of this procedure matches the well-known convergence rate of gradient descent to first-order stationary points, up to log factors. When all saddle points are non-degenerate, all second-order stationary points are local minima, and our result thus shows that perturbed gradient descent can escape saddle points almost for free. Our results can be directly applied to many machine learning applications, including deep learning. As a particular concrete example of such an application, we show that our results can be used directly to establish sharp global convergence rates for matrix factorization. Our results rely on a novel characterization of the geometry around saddle points, which may be of independent interest to the non-convex optimization community.

****

Tunable Efficient Unitary Neural Networks (EUNN) and their application to RNNs

Li Jing, Yichen Shen, Tena Dubcek, John Peurifoy, Scott Skirlo, Yann LeCun, Max Tegmark, Marin Solja■i■

Using unitary (instead of general) matrices in artificial neural networks (ANNs) is a promising way to solve the gradient explosion/vanishing problem, as well as to enable ANNs to learn long-term correlations in the data. This approach appears particularly promising for Recurrent Neural Networks (RNNs). In this work, we present a new architecture for implementing an Efficient Unitary Neural Network (EUNNs); its main advantages can be summarized as follows. Firstly, the representation capacity of the unitary space in an EUNN is fully tunable, ranging from a subspace of SU(N) to the entire unitary space. Secondly, the computational complexity for training an EUNN is merely $\mathcal{O}(1)$ per parameter. Finally, we test the performance of EUNNs on the standard copying task, the pixel-permuted MNIST digit recognition benchmark as well as the Speech Prediction Test (TIMIT). We find that our architecture significantly outperforms both other state-of-the-art unitary RNNs and the LSTM architecture, in terms of the final performance and/or the wall-clock training speed. EUNNs are thus promising alternatives to RNNs and LSTMs for a wide variety of applications.

****

An Adaptive Test of Independence with Analytic Kernel Embeddings

Wittawat Jitkrittum, Zoltán Szabó, Arthur Gretton

A new computationally efficient dependence measure, and an adaptive statistical test of independence, are proposed. The dependence measure is the difference between analytic embeddings of the joint distribution and the product of the marginals, evaluated at a finite set of locations (features). These features are chosen so as to maximize a lower bound on the test power, resulting in a test that is data-efficient, and that runs in linear time (with respect to the sample size n). The optimized features can be interpreted as evidence to reject the null hypothesis, indicating regions in the joint domain where the joint distribution and the product of the marginals differ most. Consistency of the independence test is established, for an appropriate choice of features. In real-world benchmarks, independence tests using the optimized features perform comparably to the state-of-the-art quadratic-time HSIC test, and outperform competing O(n) and O(n log n) tests.

****

StingyCD: Safely Avoiding Wasteful Updates in Coordinate Descent

Tyler B. Johnson, Carlos Guestrin

Coordinate descent (CD) is a scalable and simple algorithm for solving many optimization problems in machine learning. Despite this fact, CD can also be very computationally wasteful. Due to sparsity in sparse regression problems, for example, the majority of CD updates often result in no progress toward the solution.

To address this inefficiency, we propose a modified CD algorithm named “StingyCD.” By skipping over many updates that are guaranteed to not decrease the objective value, StingyCD significantly reduces convergence times. Since StingyCD only skips updates with this guarantee, however, StingyCD does not fully exploit the problem’s sparsity. For this reason, we also propose StingyCD+, an algorithm that achieves further speed-ups by skipping updates more aggressively. Since StingyCD and StingyCD+ rely on simple modifications to CD, it is also straightforward to use these algorithms with other approaches to scaling optimization. In empirical comparisons, StingyCD and StingyCD+ improve convergence times considerably for several L1-regularized optimization problems.

****

Differentially Private Chi-squared Test by Unit Circle Mechanism

Kazuya Kakizaki, Kazuto Fukuchi, Jun Sakuma

This paper develops differentially private mechanisms for χ^2 test of independence. While existing works put their effort into properly controlling the type-I error, in addition to that, we investigate the type-II error of differentially private mechanisms. Based on the analysis, we present unit circle mechanism: a novel differentially private mechanism based on the geometrical property of the test statistics. Compared to existing output perturbation mechanisms, our mechanism improves the dominated term of the type-II error from $O(1)$ to $O(\exp(-\sqrt{N}))$ where N is the sample size. Furthermore, we introduce novel procedures for multiple χ^2 tests by incorporating the unit circle mechanism into the sparse vector technique and the exponential mechanism. These procedures can control the family-wise error rate (FWER) properly, which has never been attained by existing mechanisms.

****

Video Pixel Networks

Nal Kalchbrenner, Aäron Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, Koray Kavukcuoglu

We propose a probabilistic video model, the Video Pixel Network (VPN), that estimates the discrete joint distribution of the raw pixel values in a video. The model and the neural architecture reflect the time, space and color structure of video tensors and encode it as a four-dimensional dependency chain. The VPN approaches the best possible performance on the Moving MNIST benchmark, a leap over the previous state of the art, and the generated videos show only minor deviations from the ground truth. The VPN also produces detailed samples on the action-conditional Robotic Pushing benchmark and generalizes to the motion of novel objects.

****

Adaptive Feature Selection: Computationally Efficient Online Sparse Linear Regression under RIP

Satyen Kale, Zohar Karnin, Tengyuan Liang, Dávid Pál

Online sparse linear regression is an online problem where an algorithm repeatedly chooses a subset of coordinates to observe in an adversarially chosen feature vector, makes a real-valued prediction, receives the true label, and incurs the squared loss. The goal is to design an online learning algorithm with sublinear regret to the best sparse linear predictor in hindsight. Without any assumptions, this problem is known to be computationally intractable. In this paper, we make the assumption that data matrix satisfies restricted isometry property, and show that this assumption leads to computationally efficient algorithms with sublinear regret for two variants of the problem. In the first variant, the true label is generated according to a sparse linear model with additive Gaussian noise. In the second, the true label is chosen adversarially.

****

Recursive Partitioning for Personalization using Observational Data

Nathan Kallus

We study the problem of learning to choose from m discrete treatment options (e.g., news item or medical drug) the one with best causal effect for a particular instance (e.g., user or patient) where the training data consists of passive observations of covariates, treatment, and the outcome of the treatment. The stan dard approach to this problem is regress and compare: split the training data by treatment, fit a regression model in each split, and, for a new instance, predict all m outcomes and pick the best. By reformulating the problem as a single learning task rather than m separate ones, we propose a new approach based on recursively partitioning the data into regimes where different treatments are optimal. We extend this approach to an optimal partitioning approach that finds a globally optimal partition, achieving a compact, interpretable, and impactful personalization model. We develop new tools for validating and evaluating personalization models on observational data and use these to demonstrate the power of our novel approaches in a personalized medicine and a job training application.

****

Multi-fidelity Bayesian Optimisation with Continuous Approximations

Kirthevasan Kandasamy, Gautam Dasarathy, Jeff Schneider, Barnabás Póczos

Bandit methods for black-box optimisation, such as Bayesian optimisation, are used in a variety of applications including hyper-parameter tuning and experiment design. Recently, multi-fidelity methods have garnered considerable attention since function evaluations have become increasingly expensive in such applications. Multi-fidelity methods use cheap approximations to the function of interest to speed up the overall optimisation process. However, most multi-fidelity methods assume only a finite number of approximations. On the other hand, in many practical applications, a continuous spectrum of approximations might be available. For instance, when tuning an expensive neural network, one might choose to approximate the cross validation performance using less data N and/or few training iterations T. Here, the approximations are best viewed as arising out of a continuous two dimensional space (N,T). In this work, we develop a Bayesian optimisation method, BOCA, for this setting. We characterise its theoretical properties and show that it achieves better regret than than strategies which ignore the approximations. BOCA outperforms several other baselines in synthetic and real experiments.

****

Schema Networks: Zero-shot Transfer with a Generative Causal Model of Intuitive Physics

Ken Kansky, Tom Silver, David A. Mély, Mohamed Eldawy, Miguel Lázaro-Gredilla, Xinghua Lou, Nimrod Dorfman, Szymon Sidor, Scott Phoenix, Dileep George

The recent adaptation of deep neural network-based methods to reinforcement learning and planning domains has yielded remarkable progress on individual tasks. Nonetheless, progress on task-to-task transfer remains limited. In pursuit of efficient and robust generalization, we introduce the Schema Network, an object-oriented generative physics simulator capable of disentangling multiple causes of events and reasoning backward through causes to achieve goals. The richly structured architecture of the Schema Network can learn the dynamics of an environment directly from data. We compare Schema Networks with Asynchronous Advantage Actor-Critic and Progressive Networks on a suite of Breakout variations, reporting results on training efficiency and zero-shot generalization, consistently demonstrating faster, more robust learning and better transfer. We argue that generalizing from limited data and learning causal relationships are essential abilities on the path toward generally intelligent systems.

****

Learning in POMDPs with Monte Carlo Tree Search

Sammie Katt, Frans A. Oliehoek, Christopher Amato

The POMDP is a powerful framework for reasoning under outcome and information uncertainty, but constructing an accurate POMDP model is difficult. Bayes-Adaptive Partially Observable Markov Decision Processes (BA-POMDPs) extend POMDPs to allow the model to be learned during execution. BA-POMDPs are a Bayesian RL approach that, in principle, allows for an optimal trade-off between exploitation and exploration. Unfortunately, BA-POMDPs are currently impractical to solve for any non-trivial domain. In this paper, we extend the Monte-Carlo Tree Search method POMCP to BA-POMDPs and show that the resulting method, which we call BA-POMCP, is able to tackle problems that previous solution methods have been unable to solve. Additionally, we introduce several techniques that exploit the BA-POMDP stru cture to improve the efficiency of BA-POMCP along with proof of their convergence.

****

Meritocratic Fairness for Cross-Population Selection

Michael Kearns, Aaron Roth, Zhiwei Steven Wu

We consider the problem of selecting a strong pool of individuals from several populations with incomparable skills (e.g. soccer players, mathematicians, and singers) in a fair manner. The quality of an individual is defined to be their relative rank (by cumulative distribution value) within their own population, which permits cross-population comparisons. We study algorithms which attempt to select the highest quality subset despite the fact that true CDF values are not known, and can only be estimated from the finite pool of candidates. Specifically, we quantify the regret in quality imposed by “meritocratic” notions of fairness, which require that individuals are selected with probability that is monotonically increasing in their true quality. We give algorithms with provable fairness and regret guarantees, as well as lower bounds, and provide empirical results which suggest that our algorithms perform better than the theory suggests. We extend our results to a sequential batch setting, in which an algorithm must repeatedly select subsets of individuals from new pools of applicants, but has the benefit of being able to compare them to the accumulated data from previous rounds.

****

On Approximation Guarantees for Greedy Low Rank Optimization

Rajiv Khanna, Ethan R. Elenberg, Alexandros G. Dimakis, Joydeep Ghosh, Sahand Negahban

We provide new approximation guarantees for greedy low rank matrix estimation under standard assumptions of restricted strong convexity and smoothness. Our novel analysis also uncovers previously unknown connections between the low rank estimation and combinatorial optimization, so much so that our bounds are reminiscent of corresponding approximation bounds in submodular maximization. Additionally, we provide also provide statistical recovery guarantees. Finally, we present empirical comparison of greedy estimation with established baselines on two important real-world problems.

****

Graph-based Isometry Invariant Representation Learning

Renata Khasanova, Pascal Frossard

Learning transformation invariant representations of visual data is an important problem in computer vision. Deep convolutional networks have demonstrated remarkable results for image and video classification tasks. However, they have achieved only limited success in the classification of images that undergo geometric transformations. In this work we present a novel Transformation Invariant Graph-based Network (TIGraNet), which learns graph-based features that are inherently invariant to isometric transformations such as rotation and translation of input images. In particular, images are represented as signals on graphs, which permits to replace classical convolution and pooling layers in deep networks with graph spectral convolution and dynamic graph pooling layers that together contribute to invariance to isometric transformation. Our experiments show high performance on rotated and translated images from the test set compared to classical architectures that are very sensitive to transformations in the data. The inherent invariance properties of our framework provide key advantages, such as increased resiliency to data variability and sustained performance with limited training sets.

****

Learning to Discover Cross-Domain Relations with Generative Adversarial Networks

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, Jiwon Kim

While humans easily recognize relations between data from different domains without any supervision, learning to automatically discover them is in general very challenging and needs many ground-truth pairs that illustrate the relations. To avoid costly pairing, we address the task of discovering cross-domain relations given unpaired data. We propose a method based on a generative adversarial network that learns to discover relations between different domains (DiscoGAN). Using the discovered relations, our proposed network successfully transfers style from one domain to another while preserving key attributes such as orientation and face identity.

****

SplitNet: Learning to Semantically Split Deep Networks for Parameter Reduction and Model Parallelization

Juyong Kim, Yookoon Park, Gunhee Kim, Sung Ju Hwang

We propose a novel deep neural network that is both lightweight and effectively structured for model parallelization. Our network, which we name as SplitNet, automatically learns to split the network weights into either a set or a hierarchy of multiple groups that use disjoint sets of features, by learning both the class-to-group and feature-to-group assignment matrices along with the network weights. This produces a tree-structured network that involves no connection between branched subtrees of semantically disparate class groups. SplitNet thus greatly reduces the number of parameters and requires significantly less computations, and is also embarrassingly model parallelizable at test time, since the network evaluation for each subnetwork is completely independent except for the shared lower layer weights that can be duplicated over multiple processors. We validate our method with two deep network models (ResNet and AlexNet) on two different datasets (CIFAR-100 and ILSVRC 2012) for image classification, on which our method obtains networks with significantly reduced number of parameters while achieving comparable or superior classification accuracies over original full deep networks, and accelerated test speed with multiple GPUs.

****

Cost-Optimal Learning of Causal Graphs

Murat Kocaoglu, Alex Dimakis, Sriram Vishwanath

We consider the problem of learning a causal graph over a set of variables with interventions. We study the cost-optimal causal graph learning problem: For a given skeleton (undirected version of the causal graph), design the set of interventions with minimum total cost, that can uniquely identify any causal graph with the given skeleton. We show that this problem is solvable in polynomial time. Later, we consider the case when the number of interventions is limited. For this case, we provide polynomial time algorithms when the skeleton is a tree or a clique tree. For a general chordal skeleton, we develop an efficient greedy algorithm, which can be improved when the causal graph skeleton is an interval graph.

****

Understanding Black-box Predictions via Influence Functions

Pang Wei Koh, Percy Liang

How can we explain the predictions of a black-box model? In this paper, we use influence functions — a classic technique from robust statistics — to trace a model’s prediction through the learning algorithm and back to its training data, thereby identifying training points most responsible for a given prediction. To scale up influence functions to modern machine learning settings, we develop a simple, efficient implementation that requires only oracle access to gradients and Hessian-vector products. We show that even on non-convex and non-differentiable models where the theory breaks down, approximations to influence functions can still provide valuable information. On linear models and convolutional neural networks, we demonstrate that influence functions are useful for multiple purposes: understanding model behavior, debugging models, detecting dataset errors, and even creating visually-indistinguishable training-set attacks.

****

Sub-sampled Cubic Regularization for Non-convex Optimization

Jonas Moritz Kohler, Aurelien Lucchi

We consider the minimization of non-convex functions that typically arise in machine learning. Specifically, we focus our attention on a variant of trust region methods known as cubic regularization. This approach is particularly attractive because it escapes strict saddle points and it provides stronger convergence guarantees than first- and second-order as well as classical trust region methods. However, it suffers from a high computational complexity that makes it impractical for large-scale learning. Here, we propose a novel method that uses sub-samp ling to lower this computational cost. By the use of concentration inequalities we provide a sampling scheme that gives sufficiently accurate gradient and Hessian approximations to retain the strong global and local convergence guarantees of cubically regularized methods. To the best of our knowledge this is the first work that gives global convergence guarantees for a sub-sampled variant of cubic regularization on non-convex functions. Furthermore, we provide experimental results supporting our theory.

****

PixelCNN Models with Auxiliary Variables for Natural Image Modeling

Alexander Kolesnikov, Christoph H. Lampert

We study probabilistic models of natural images and extend the autoregressive family of PixelCNN models by incorporating auxiliary variables. Subsequently, we describe two new generative image models that exploit different image transformations as auxiliary variables: a quantized grayscale view of the image or a multi-resolution image pyramid. The proposed models tackle two known shortcomings of existing PixelCNN models: 1) their tendency to focus on low-level image details, while largely ignoring high-level image information, such as object shapes, and 2) their computationally costly procedure for image sampling. We experimentally demonstrate benefits of our models, in particular showing that they produce much more realistically looking image samples than previous state-of-the-art probabilistic models.

****

Active Learning for Cost-Sensitive Classification

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, John Langford

We design an active learning algorithm for cost-sensitive multiclass classification: problems where different errors have different costs. Our algorithm, COAL, makes predictions by regressing to each label’s cost and predicting the smallest. On a new example, it uses a set of regressors that perform well on past data to estimate possible costs for each label. It queries only the labels that could be the best, ignoring the sure losers. We prove COAL can be efficiently implemented for any regression family that admits squared loss optimization; it also enjoys strong guarantees with respect to predictive performance and labeling effort. Our experiment with COAL show significant improvements in labeling effort and test cost over passive and active baselines.

****

Evaluating Bayesian Models with Posterior Dispersion Indices

Alp Kucukelbir, Yixin Wang, David M. Blei

Probabilistic modeling is cyclical: we specify a model, infer its posterior, and evaluate its performance. Evaluation drives the cycle, as we revise our model based on how it performs. This requires a metric. Traditionally, predictive accuracy prevails. Yet, predictive accuracy does not tell the whole story. We propose to evaluate a model through posterior dispersion. The idea is to analyze how each datapoint fares in relation to posterior uncertainty around the hidden structure. This highlights datapoints the model struggles to explain and provides complimentary insight to datapoints with low predictive accuracy. We present a family of posterior dispersion indices (PDI) that capture this idea. We show how a PDI identifies patterns of model mismatch in three real data examples: voting preferences, supermarket shopping, and population genetics.

****

Resource-efficient Machine Learning in 2 KB RAM for the Internet of Things

Ashish Kumar, Saurabh Goyal, Manik Varma

This paper develops a novel tree-based algorithm, called Bonsai, for efficient prediction on IoT devices – such as those based on the Arduino Uno board having an 8 bit ATmega328P microcontroller operating at 16 MHz with no native floating point support, 2 KB RAM and 32 KB read-only flash. Bonsai maintains prediction accuracy while minimizing model size and prediction costs by: (a) developing a tree model which learns a single, shallow, sparse tree with powerful nodes; (b) sparsely projecting all data into a low-dimensional space in which the tree is learnt; and (c) jointly learning all tree and projection parameters. Experimental results on multiple benchmark datasets demonstrate that Bonsai can make prediction s in milliseconds even on slow microcontrollers, can fit in KB of memory, has lower battery consumption than all other algorithms while achieving prediction accuracies that can be as much as 30\% higher than state-of-the-art methods for resource-efficient machine learning. Bonsai is also shown to generalize to other resource constrained settings beyond IoT by generating significantly better search results as compared to Bing’s L3 ranker when the model size is restricted to 300 bytes. Bonsai’s code can be downloaded from (http://www.manikvarma.org/code/Bonsai/download.html).

****

Grammar Variational Autoencoder
Matt J. Kusner, Brooks Paige, José Miguel Hernández-Lobato
Deep generative models have been wildly successful at learning coherent latent representations for continuous data such as natural images, artwork, and audio. However, generative modeling of discrete data such as arithmetic expressions and molecular structures still poses significant challenges. Crucially, state-of-the-art methods often produce outputs that are not valid. We make the key observation that frequently, discrete data can be represented as a parse tree from a context-free grammar. We propose a variational autoencoder which directly encodes from and decodes to these parse trees, ensuring the generated outputs are always syntactically valid. Surprisingly, we show that not only does our model more often generate valid outputs, it also learns a more coherent latent space in which nearby points decode to similar discrete outputs. We demonstrate the effectiveness of our learned models by showing their improved performance in Bayesian optimization for symbolic regression and molecule generation.

****

Co-clustering through Optimal Transport
Charlotte Laclau, Ievgen Redko, Basarab Matei, Younès Bennani, Vincent Brault
In this paper, we present a novel method for co-clustering, an unsupervised learning approach that aims at discovering homogeneous groups of data instances and features by grouping them simultaneously. The proposed method uses the entropy regularized optimal transport between empirical measures defined on data instances and features in order to obtain an estimated joint probability density function represented by the optimal coupling matrix. This matrix is further factorized to obtain the induced row and columns partitions using multiscale representations approach. To justify our method theoretically, we show how the solution of the regularized optimal transport can be seen from the variational inference perspective thus motivating its use for co-clustering. The algorithm derived for the proposed method and its kernelized version based on the notion of Gromov-Wasserstein distance are fast, accurate and can determine automatically the number of both row and column clusters. These features are vividly demonstrated through extensive experimental evaluations.

****

Conditional Accelerated Lazy Stochastic Gradient Descent
Guanghui Lan, Sebastian Pokutta, Yi Zhou, Daniel Zink
In this work we introduce a conditional accelerated lazy stochastic gradient descent algorithm with optimal number of calls to a stochastic first-order oracle and convergence rate $O(1/\epsilon^2)$ improving over the projection-free, Online Frank-Wolfe based stochastic gradient descent of (Hazan and Kale, 2012) with convergence rate $O(1/\epsilon^4)$.

****

Consistent k-Clustering
Silvio Lattanzi, Sergei Vassilvitskii
The study of online algorithms and competitive analysis provides a solid foundation for studying the quality of irrevocable decision making when the data arrives in an online manner. While in some scenarios the decisions are indeed irrevocable, there are many practical situations when changing a previous decision is not impossible, but simply expensive. In this work we formalize this notion and introduce the consistent k-clustering problem. With points arriving online, the goal is to maintain a constant approximate solution, while minimizing the number of reclusterings necessary. We prove a lower bound, showing that \$\Omega(k \log n

)$ changes are necessary in the worst case for a wide range of objective functions. On the positive side, we give an algorithm that needs only $O(k^2 \log^4n)$ changes to maintain a constant competitive solution, an exponential improvement on the naive solution of reclustering at every time step. Finally, we show experimentally that our approach performs much better than the theoretical bound, with the number of changes growing approximately as $O(\log n)$.

****

Deep Spectral Clustering Learning

Marc T. Law, Raquel Urtasun, Richard S. Zemel

Clustering is the task of grouping a set of examples so that similar examples are grouped into the same cluster while dissimilar examples are in different clusters. The quality of a clustering depends on two problem-dependent factors which are i) the chosen similarity metric and ii) the data representation. Supervised clustering approaches, which exploit labeled partitioned datasets have thus been proposed, for instance to learn a metric optimized to perform clustering. However, most of these approaches assume that the representation of the data is fixed and then learn an appropriate linear transformation. Some deep supervised clustering learning approaches have also been proposed. However, they rely on iterative methods to compute gradients resulting in high algorithmic complexity. In this paper, we propose a deep supervised clustering metric learning method that formulates a novel loss function. We derive a closed-form expression for the gradient that is efficient to compute: the complexity to compute the gradient is linear in the size of the training mini-batch and quadratic in the representation dimensionality. We further reveal how our approach can be seen as learning spectral clustering. Experiments on standard real-world datasets confirm state-of-the-art Recall@K performance.

****

Coordinated Multi-Agent Imitation Learning

Hoang M. Le, Yisong Yue, Peter Carr, Patrick Lucey

We study the problem of imitation learning from demonstrations of multiple coordinating agents. One key challenge in this setting is that learning a good model of coordination can be difficult, since coordination is often implicit in the demonstrations and must be inferred as a latent variable. We propose a joint approach that simultaneously learns a latent coordination model along with the individual policies. In particular, our method integrates unsupervised structure learning with conventional imitation learning. We illustrate the power of our approach on a difficult problem of learning multiple policies for fine-grained behavior modeling in team sports, where different players occupy different roles in the coordinated team strategy. We show that having a coordination model to infer the roles of players yields substantially improved imitation loss compared to conventional baselines.

****

Bayesian inference on random simple graphs with power law degree distributions

Juho Lee, Creighton Heaukulani, Zoubin Ghahramani, Lancelot F. James, Seungjin Choi

We present a model for random simple graphs with power law (i.e., heavy-tailed) degree distributions. To attain this behavior, the edge probabilities in the graph are constructed from Bertoin–Fujita–Roynette–Yor (BFRY) random variables, which have been recently utilized in Bayesian statistics for the construction of power law models in several applications. Our construction readily extends to capture the structure of latent factors, similarly to stochastic block-models, while maintaining its power law degree distribution. The BFRY random variables are well approximated by gamma random variables in a variational Bayesian inference routine, which we apply to several network datasets for which power law degree distributions are a natural assumption. By learning the parameters of the BFRY distribution via probabilistic inference, we are able to automatically select the appropriate power law behavior from the data. In order to further scale our inference procedure, we adopt stochastic gradient ascent routines where the gradients are computed on minibatches (i.e., subsets) of the edges in the graph.

****

Confident Multiple Choice Learning

Kimin Lee, Changho Hwang, KyoungSoo Park, Jinwoo Shin

Ensemble methods are arguably the most trustworthy techniques for boosting the performance of machine learning models. Popular independent ensembles (IE) relying on naive averaging/voting scheme have been of typical choice for most applications involving deep neural networks, but they do not consider advanced collaboration among ensemble models. In this paper, we propose new ensemble methods specialized for deep neural networks, called confident multiple choice learning (CMCL): it is a variant of multiple choice learning (MCL) via addressing its overconfidence issue.In particular, the proposed major components of CMCL beyond the original MCL scheme are (i) new loss, i.e., confident oracle loss, (ii) new architecture, i.e., feature sharing and (iii) new training method, i.e., stochastic labeling. We demonstrate the effect of CMCL via experiments on the image classification on CIFAR and SVHN, and the foreground-background segmentation on the iCoseg. In particular, CMCL using 5 residual networks provides 14.05\% and 6.60\% relative reductions in the top-1 error rates from the corresponding IE scheme for the classification task on CIFAR and SVHN, respectively.

****

Deriving Neural Architectures from Sequence and Graph Kernels

Tao Lei, Wengong Jin, Regina Barzilay, Tommi Jaakkola

The design of neural architectures for structured objects is typically guided by experimental insights rather than a formal process. In this work, we appeal to kernels over combinatorial structures, such as sequences and graphs, to derive appropriate neural operations. We introduce a class of deep recurrent neural operations and formally characterize their associated kernel spaces. Our recurrent modules compare the input to virtual reference objects (cf. filters in CNN) via the kernels. Similar to traditional neural operations, these reference objects are parameterized and directly optimized in end-to-end training. We empirically evaluate the proposed class of neural architectures on standard applications such as language modeling and molecular graph regression, achieving state-of-the-art results across these applications.

****

Doubly Greedy Primal-Dual Coordinate Descent for Sparse Empirical Risk Minimization

Qi Lei, Ian En-Hsu Yen, Chao-yuan Wu, Inderjit S. Dhillon, Pradeep Ravikumar

We consider the popular problem of sparse empirical risk minimization with linear predictors and a large number of both features and observations. With a convex-concave saddle point objective reformulation, we propose a Doubly Greedy Primal-Dual Coordinate Descent algorithm that is able to exploit sparsity in both primal and dual variables. It enjoys a low cost per iteration and our theoretical analysis shows that it converges linearly with a good iteration complexity, provided that the set of primal variables is sparse. We then extend this algorithm further to leverage active sets. The resulting new algorithm is even faster, and experiments on large-scale Multi-class data sets show that our algorithm achieves up to 30 times speedup on several state-of-the-art optimization methods.

****

Learning to Align the Source Code to the Compiled Object Code

Dor Levy, Lior Wolf

We propose a new neural network architecture and use it for the task of statement-by-statement alignment of source code and its compiled object code. Our architecture learns the alignment between the two sequences – one being the translation of the other – by mapping each statement to a context-dependent representation vector and aligning such vectors using a grid of the two sequence domains. Our experiments include short C functions, both artificial and human-written, and show that our neural network architecture is able to predict the alignment with high accuracy, outperforming known baselines. We also demonstrate that our model is general and can learn to solve graph problems such as the Traveling Salesman Problem.

****

Dropout Inference in Bayesian Neural Networks with Alpha-divergences

Yingzhen Li, Yarin Gal

To obtain uncertainty estimates with real-world Bayesian deep learning models, practical inference approximations are needed. Dropout variational inference (VI) for example has been used for machine vision and medical applications, but VI can severely underestimates model uncertainty. Alpha-divergences are alternative divergences to VI’s KL objective, which are able to avoid VI’s uncertainty underestimation. But these are hard to use in practice: existing techniques can only use Gaussian approximating distributions, and require existing models to be changed radically, thus are of limited use for practitioners. We propose a re-parametrisation of the alpha-divergence objectives, deriving a simple inference technique which, together with dropout, can be easily implemented with existing models by simply changing the loss of the model. We demonstrate improved uncertainty estimates and accuracy compared to VI in dropout networks. We study our model’s epistemic uncertainty far away from the data using adversarial images, showing that these can be distinguished from non-adversarial images by examining our model’s uncertainty.

****

Provable Alternating Gradient Descent for Non-negative Matrix Factorization with Strong Correlations

Yuanzhi Li, Yingyu Liang

Non-negative matrix factorization is a basic tool for decomposing data into the feature and weight matrices under non-negativity constraints, and in practice is often solved in the alternating minimization framework. However, it is unclear whether such algorithms can recover the ground-truth feature matrix when the weights for different features are highly correlated, which is common in applications. This paper proposes a simple and natural alternating gradient descent based algorithm, and shows that with a mild initialization it provably recovers the ground-truth in the presence of strong correlations. In most interesting cases, the correlation can be in the same order as the highest possible. Our analysis also reveals its several favorable features including robustness to noise. We complement our theoretical results with empirical studies on semi-synthetic datasets, demonstrating its advantage over several popular methods in recovering the ground-truth.

****

Provably Optimal Algorithms for Generalized Linear Contextual Bandits

Lihong Li, Yu Lu, Dengyong Zhou

Contextual bandits are widely used in Internet services from news recommendation to advertising, and to Web search. Generalized linear models (logistical regression in particular) have demonstrated stronger performance than linear models in many applications where rewards are binary. However, most theoretical analyses on contextual bandits so far are on linear bandits. In this work, we propose an upper confidence bound based algorithm for generalized linear contextual bandits, which achieves an $\sim O(\sqrt{dT})$ regret over T rounds with d dimensional feature vectors. This regret matches the minimax lower bound, up to logarithmic terms, and improves on the best previous result by a $\sqrt{d}$ factor, assuming the number of arms is fixed. A key component in our analysis is to establish a new, sharp finite-sample confidence bound for maximum likelihood estimates in generalized linear models, which may be of independent interest. We also analyze a simpler upper confidence bound algorithm, which is useful in practice, and prove it to have optimal regret for certain cases.

****

Fast k-Nearest Neighbour Search via Prioritized DCI

Ke Li, Jitendra Malik

Most exact methods for k-nearest neighbour search suffer from the curse of dimensionality; that is, their query times exhibit exponential dependence on either the ambient or the intrinsic dimensionality. Dynamic Continuous Indexing (DCI) offers a promising way of circumventing the curse and successfully reduces the dependence of query time on intrinsic dimensionality from exponential to sublinear. In this paper, we propose a variant of DCI, which we call Prioritized DCI, and show a remarkable improvement in the dependence of query time on intrinsic dimen sionality. In particular, a linear increase in intrinsic dimensionality, or equivalently, an exponential increase in the number of points near a query, can be mostly counteracted with just a linear increase in space. We also demonstrate empirically that Prioritized DCI significantly outperforms prior methods. In particular, relative to Locality-Sensitive Hashing (LSH), Prioritized DCI reduces the number of distance evaluations by a factor of 14 to 116 and the memory consumption by a factor of 21.

****

Forest-type Regression with General Losses and Robust Forest
Alexander Hanbo Li, Andrew Martin

This paper introduces a new general framework for forest-type regression which allows the development of robust forest regressors by selecting from a large family of robust loss functions. In particular, when plugged in the squared error and quantile losses, it will recover the classical random forest and quantile random forest. We then use robust loss functions to develop more robust forest-type regression algorithms. In the experiments, we show by simulation and real data that our robust forests are indeed much more insensitive to outliers, and choosing the right number of nearest neighbors can quickly improve the generalization performance of random forest.

****

Stochastic Modified Equations and Adaptive Stochastic Gradient Algorithms
Qianxiao Li, Cheng Tai, Weinan E

We develop the method of stochastic modified equations (SME), in which stochastic gradient algorithms are approximated in the weak sense by continuous-time stochastic differential equations. We exploit the continuous formulation together with optimal control theory to derive novel adaptive hyper-parameter adjustment policies. Our algorithms have competitive performance with the added benefit of being robust to varying models and datasets. This provides a general methodology for the analysis and design of stochastic gradient algorithms.

****

Convergence Analysis of Proximal Gradient with Momentum for Nonconvex Optimization
Qunwei Li, Yi Zhou, Yingbin Liang, Pramod K. Varshney

In this work, we investigate the accelerated proximal gradient method for nonconvex programming (APGnc). The method compares between a usual proximal gradient step and a linear extrapolation step, and accepts the one that has a lower function value to achieve a monotonic decrease. In specific, under a general nonsmooth and nonconvex setting, we provide a rigorous argument to show that the limit points of the sequence generated by APGnc are critical points of the objective function. Then, by exploiting the Kurdyka-Lojasiewicz (KL) property for a broad class of functions, we establish the linear and sub-linear convergence rates of the function value sequence generated by APGnc. We further propose a stochastic variance reduced APGnc (SVRG-APGnc), and establish its linear convergence under a special case of the KL property. We also extend the analysis to the inexact version of these methods and develop an adaptive momentum strategy that improves the numerical performance.

****

Exact MAP Inference by Avoiding Fractional Vertices
Erik M. Lindgren, Alexandros G. Dimakis, Adam Klivans

Given a graphical model, one essential problem is MAP inference, that is, finding the most likely configuration of states according to the model. Although this problem is NP-hard, large instances can be solved in practice and it is a major open question is to explain why this is true. We give a natural condition under which we can provably perform MAP inference in polynomial time—we require that the number of fractional vertices in the LP relaxation exceeding the optimal solution is bounded by a polynomial in the problem size. This resolves an open question by Dimakis, Gohari, and Wainwright. In contrast, for general LP relaxations of integer programs, known techniques can only handle a constant number of fractional vertices whose value exceeds the optimal solution. We experimentally verify this condition and demonstrate how efficient various integer programming metho ds are at removing fractional solutions.

****

Leveraging Union of Subspace Structure to Improve Constrained Clustering
John Lipor, Laura Balzano

Many clustering problems in computer vision and other contexts are also classification problems, where each cluster shares a meaningful label. Subspace clustering algorithms in particular are often applied to problems that fit this description, for example with face images or handwritten digits. While it is straightforward to request human input on these datasets, our goal is to reduce this input as much as possible. We present a pairwise-constrained clustering algorithm that actively selects queries based on the union-of-subspaces model. The central step of the algorithm is in querying points of minimum margin between estimated subspaces; analogous to classifier margin, these lie near the decision boundary. We prove that points lying near the intersection of subspaces are points with low margin. Our procedure can be used after any subspace clustering algorithm that outputs an affinity matrix. We demonstrate on several datasets that our algorithm drives the clustering error down considerably faster than the state-of-the-art active query algorithms on datasets with subspace structure and is competitive on other datasets.

****

Zero-Inflated Exponential Family Embeddings
Li-Ping Liu, David M. Blei

Word embeddings are a widely-used tool to analyze language, and exponential family embeddings (Rudolph et al., 2016) generalize the technique to other types of data. One challenge to fitting embedding methods is sparse data, such as a document/term matrix that contains many zeros. To address this issue, practitioners typically downweight or subsample the zeros, thus focusing learning on the non-zero entries. In this paper, we develop zero-inflated embeddings, a new embedding method that is designed to learn from sparse observations. In a zero-inflated embedding (ZIE), a zero in the data can come from an interaction to other data (i.e., an embedding) or from a separate process by which many observations are equal to zero (i.e. a probability mass at zero). Fitting a ZIE naturally downweights the zeros and dampens their influence on the model. Across many types of data—language, movie ratings, shopping histories, and bird watching logs—we found that zero-inflated embeddings provide improved predictive performance over standard approaches and find better vector representation of items.

****

Iterative Machine Teaching
Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B. Smith, James M. Rehg, Le Song

In this paper, we consider the problem of machine teaching, the inverse problem of machine learning. Different from traditional machine teaching which views the learners as batch algorithms, we study a new paradigm where the learner uses an iterative algorithm and a teacher can feed examples sequentially and intelligently based on the current performance of the learner. We show that the teaching complexity in the iterative case is very different from that in the batch case. Instead of constructing a minimal training set for learners, our iterative machine teaching focuses on achieving fast convergence in the learner model. Depending on the level of information the teacher has from the learner model, we design teaching algorithms which can provably reduce the number of teaching examples and achieve faster convergence than learning without teachers. We also validate our theoretical findings with extensive experiments on different data distribution and real image datasets.

****

Algorithmic Stability and Hypothesis Complexity
Tongliang Liu, Gábor Lugosi, Gergely Neu, Dacheng Tao

We introduce a notion of algorithmic stability of learning algorithms—that we term hypothesis stability—that captures stability of the hypothesis output by the learning algorithm in the normed space of functions from which hypotheses are selected. The main result of the paper bounds the generalization error of any lear ning algorithm in terms of its hypothesis stability. The bounds are based on martingale inequalities in the Banach space to which the hypotheses belong. We apply the general bounds to bound the performance of some learning algorithms based on empirical risk minimization and stochastic gradient descent.

****

Analogical Inference for Multi-relational Embeddings

Hanxiao Liu, Yuexin Wu, Yiming Yang

Large-scale multi-relational embedding refers to the task of learning the latent representations for entities and relations in large knowledge graphs. An effective and scalable solution for this problem is crucial for the true success of knowledge-based inference in a broad range of applications. This paper proposes a novel framework for optimizing the latent representations with respect to the analogical properties of the embedded entities and relations. By formulating the objective function in a differentiable fashion, our model enjoys both its theoretical power and computational scalability, and significantly outperformed a large number of representative baseline methods on benchmark datasets. Furthermore, the model offers an elegant unification of several well-known methods in multi-relational embedding, which can be proven to be special instantiations of our framework.

****

Dual Iterative Hard Thresholding: From Non-convex Sparse Minimization to Non-smooth Concave Maximization

Bo Liu, Xiao-Tong Yuan, Lezi Wang, Qingshan Liu, Dimitris N. Metaxas

Iterative Hard Thresholding (IHT) is a class of projected gradient descent methods for optimizing sparsity-constrained minimization models, with the best known efficiency and scalability in practice. As far as we know, the existing IHT-style methods are designed for sparse minimization in primal form. It remains open to explore duality theory and algorithms in such a non-convex and NP-hard setting. In this article, we bridge the gap by establishing a duality theory for sparsity-constrained minimization with ℓ_2-regularized objective and proposing an IHT-style algorithm for dual maximization. Our sparse duality theory provides a set of sufficient and necessary conditions under which the original NP-hard/non-convex problem can be equivalently solved in a dual space. The proposed dual IHT algorithm is a super-gradient method for maximizing the non-smooth dual objective. An interesting finding is that the sparse recovery performance of dual IHT is invariant to the Restricted Isometry Property (RIP), which is required by all the existing primal IHT without sparsity relaxation. Moreover, a stochastic variant of dual IHT is proposed for large-scale stochastic optimization. Numerical results demonstrate that dual IHT algorithms can achieve more accurate model estimation given small number of training data and have higher computational efficiency than the state-of-the-art primal IHT-style algorithms.

****

Gram-CTC: Automatic Unit Selection and Target Decomposition for Sequence Labelling

Hairong Liu, Zhenyao Zhu, Xiangang Li, Sanjeev Satheesh

Most existing sequence labelling models rely on a fixed decomposition of a target sequence into a sequence of basic units. These methods suffer from two major drawbacks: 1) the set of basic units is fixed, such as the set of words, characters or phonemes in speech recognition, and 2) the decomposition of target sequences is fixed. These drawbacks usually result in sub-optimal performance of modeling sequences. In this paper, we extend the popular CTC loss criterion to alleviate these limitations, and propose a new loss function called Gram-CTC. While preserving the advantages of CTC, Gram-CTC automatically learns the best set of basic units (grams), as well as the most suitable decomposition of target sequences. Unlike CTC, Gram-CTC allows the model to output variable number of characters at each time step, which enables the model to capture longer term dependency and improves the computational efficiency. We demonstrate that the proposed Gram-CTC improves CTC in terms of both performance and efficiency on the large vocabulary speech recognition task at multiple scales of data, and that with Gram-CTC we can outperform the state-of-the-art on a standard speech benchmark.

****

Learning Infinite Layer Networks Without the Kernel Trick
Roi Livni, Daniel Carmon, Amir Globerson

Infinite Layer Networks (ILN) have been proposed as an architecture that mimics neural networks while enjoying some of the advantages of kernel methods. ILN are networks that integrate over infinitely many nodes within a single hidden layer. It has been demonstrated by several authors that the problem of learning ILN can be reduced to the kernel trick, implying that whenever a certain integral can be computed analytically they are efficiently learnable. In this work we give an online algorithm for ILN, which avoids the kernel trick assumption. More generally and of independent interest, we show that kernel methods in general can be exploited even when the kernel cannot be efficiently computed but can only be estimated via sampling. We provide a regret analysis for our algorithm, showing that it matches the sample complexity of methods which have access to kernel values. Thus, our method is the first to demonstrate that the kernel trick is not necessary, as such, and random features suffice to obtain comparable performance.

****

Deep Transfer Learning with Joint Adaptation Networks
Mingsheng Long, Han Zhu, Jianmin Wang, Michael I. Jordan

Deep networks have been successfully applied to learn transferable features for adapting models from a source domain to a different target domain. In this paper, we present joint adaptation networks (JAN), which learn a transfer network by aligning the joint distributions of multiple domain-specific layers across domains based on a joint maximum mean discrepancy (JMMD) criterion. Adversarial training strategy is adopted to maximize JMMD such that the distributions of the source and target domains are made more distinguishable. Learning can be performed by stochastic gradient descent with the gradients computed by back-propagation in linear-time. Experiments testify that our model yields state of the art results on standard datasets.

****

Multiplicative Normalizing Flows for Variational Bayesian Neural Networks
Christos Louizos, Max Welling

We reinterpret multiplicative noise in neural networks as auxiliary random variables that augment the approximate posterior in a variational setting for Bayesian neural networks. We show that through this interpretation it is both efficient and straightforward to improve the approximation by employing normalizing flows while still allowing for local reparametrizations and a tractable lower bound. In experiments we show that with this new approximation we can significantly improve upon classical mean field for Bayesian neural networks on both predictive accuracy as well as predictive uncertainty.

****

How Close Are the Eigenvectors of the Sample and Actual Covariance Matrices?
Andreas Loukas

How many samples are sufficient to guarantee that the eigenvectors of the sample covariance matrix are close to those of the actual covariance matrix? For a wide family of distributions, including distributions with finite second moment and sub-gaussian distributions supported in a centered Euclidean ball, we prove that the inner product between eigenvectors of the sample and actual covariance matrices decreases proportionally to the respective eigenvalue distance and the number of samples. Our findings imply non-asymptotic concentration bounds for eigenvectors and eigenvalues and carry strong consequences for the non-asymptotic analysis of PCA and its applications. For instance, they provide conditions for separating components estimated from $O(1)$ samples and show that even few samples can be sufficient to perform dimensionality reduction, especially for low-rank covariances.

****

Learning Deep Architectures via Generalized Whitened Neural Networks
Ping Luo

Whitened Neural Network (WNN) is a recent advanced deep architecture, which improves convergence and generalization of canonical neural networks by whitening th eir internal hidden representation. However, the whitening transformation increases computation time. Unlike WNN that reduced runtime by performing whitening every thousand iterations, which degenerates convergence due to the ill conditioning, we present generalized WNN (GWNN), which has three appealing properties. First, GWNN is able to learn compact representation to reduce computations. Second, it enables whitening transformation to be performed in a short period, preserving good conditioning. Third, we propose a data-independent estimation of the covariance matrix to further improve computational efficiency. Extensive experiments on various datasets demonstrate the benefits of GWNN.

****

Learning Gradient Descent: Better Generalization and Longer Horizons

Kaifeng Lv, Shunhua Jiang, Jian Li

Training deep neural networks is a highly nontrivial task, involving carefully selecting appropriate training algorithms, scheduling step sizes and tuning other hyperparameters. Trying different combinations can be quite labor-intensive and time consuming. Recently, researchers have tried to use deep learning algorithms to exploit the landscape of the loss function of the training problem of interest, and learn how to optimize over it in an automatic way. In this paper, we propose a new learning-to-learn model and some useful and practical tricks. Our optimizer outperforms generic, hand-crafted optimization algorithms and state-of-the-art learning-to-learn optimizers by DeepMind in many tasks. We demonstrate the effectiveness of our algorithms on a number of tasks, including deep MLPs, CNNs, and simple LSTMs.

****

Spherical Structured Feature Maps for Kernel Approximation

Yueming Lyu

We propose Spherical Structured Feature (SSF) maps to approximate shift and rotation invariant kernels as well as b^{th}-order arc-cosine kernels (Cho \& Saul, 2009). We construct SSF maps based on the point set on $d-1$ dimensional sphere $\mathbb{S}^{d-1}$. We prove that the inner product of SSF maps are unbiased estimates for above kernels if asymptotically uniformly distributed point set on $\mathbb{S}^{d-1}$ is given. According to (Brauchart \& Grabner, 2015), optimizing the discrete Riesz s-energy can generate asymptotically uniformly distributed point set on $\mathbb{S}^{d-1}$. Thus, we propose an efficient coordinate decent method to find a local optimum of the discrete Riesz s-energy for SSF maps construction. Theoretically, SSF maps construction achieves linear space complexity and loglinear time complexity. Empirically, SSF maps achieve superior performance compared with other methods.

****

Stochastic Gradient MCMC Methods for Hidden Markov Models

Yi-An Ma, Nicholas J. Foti, Emily B. Fox

Stochastic gradient MCMC (SG-MCMC) algorithms have proven useful in scaling Bayesian inference to large datasets under an assumption of i.i.d data. We instead develop an SG-MCMC algorithm to learn the parameters of hidden Markov models (HMMs) for time-dependent data. There are two challenges to applying SG-MCMC in this setting: The latent discrete states, and needing to break dependencies when considering minibatches. We consider a marginal likelihood representation of the HMM and propose an algorithm that harnesses the inherent memory decay of the process. We demonstrate the effectiveness of our algorithm on synthetic experiments and an ion channel recording data, with runtimes significantly outperforming batch MCMC.

****

Self-Paced Co-training

Fan Ma, Deyu Meng, Qi Xie, Zina Li, Xuanyi Dong

Co-training is a well-known semi-supervised learning approach which trains classifiers on two different views and exchanges labels of unlabeled instances in an iterative way. During co-training process, labels of unlabeled instances in the training pool are very likely to be false especially in the initial training rounds, while the standard co-training algorithm utilizes a “draw without replacement” manner and does not remove these false labeled instances from training. This issue not only tends to degenerate its performance but also hampers its fundamental theory. Besides, there is no optimization model to explain what objective a cotraining process optimizes. To these issues, in this study we design a new co-training algorithm named self-paced cotraining (SPaCo) with a “draw with replacement” learning mode. The rationality of SPaCo can be proved under theoretical assumptions utilized in traditional co-training research, and furthermore, the algorithm exactly complies with the alternative optimization process for an optimization model of self-paced curriculum learning, which can be finely explained in robust learning manner. Experimental results substantiate the superiority of the proposed method as compared with current state-of-the-art co-training methods.

****

Interactive Learning from Policy-Dependent Human Feedback

James MacGlashan, Mark K. Ho, Robert Loftin, Bei Peng, Guan Wang, David L. Roberts, Matthew E. Taylor, Michael L. Littman

This paper investigates the problem of interactively learning behaviors communicated by a human teacher using positive and negative feedback. Much previous work on this problem has made the assumption that people provide feedback for decisions that is dependent on the behavior they are teaching and is independent from the learner’s current policy. We present empirical results that show this assumption to be false—whether human trainers give a positive or negative feedback for a decision is influenced by the learner’s current policy. Based on this insight, we introduce Convergent Actor-Critic by Humans (COACH), an algorithm for learning from policy-dependent feedback that converges to a local optimum. Finally, we demonstrate that COACH can successfully learn multiple behaviors on a physical robot.

****

A Laplacian Framework for Option Discovery in Reinforcement Learning

Marlos C. Machado, Marc G. Bellemare, Michael Bowling

Representation learning and option discovery are two of the biggest challenges in reinforcement learning (RL). Proto-value functions (PVFs) are a well-known approach for representation learning in MDPs. In this paper we address the option discovery problem by showing how PVFs implicitly define options. We do it by introducing eigenpurposes, intrinsic reward functions derived from the learned representations. The options discovered from eigenpurposes traverse the principal directions of the state space. They are useful for multiple tasks because they are discovered without taking the environment’s rewards into consideration. Moreover, different options act at different time scales, making them helpful for exploration. We demonstrate features of eigenpurposes in traditional tabular domains as well as in Atari 2600 games.

****

Frame-based Data Factorizations

Sebastian Mair, Ahcène Boubekki, Ulf Brefeld

Archetypal Analysis is the method of choice to compute interpretable matrix factorizations. Every data point is represented as a convex combination of factors, i.e., points on the boundary of the convex hull of the data. This renders computation inefficient. In this paper, we show that the set of vertices of a convex hull, the so-called frame, can be efficiently computed by a quadratic program. We provide theoretical and empirical results for our proposed approach and make use of the frame to accelerate Archetypal Analysis. The novel method yields similar reconstruction errors as baseline competitors but is much faster to compute.

****

Global optimization of Lipschitz functions

Cédric Malherbe, Nicolas Vayatis

The goal of the paper is to design sequential strategies which lead to efficient optimization of an unknown function under the only assumption that it has a finite Lipschitz constant. We first identify sufficient conditions for the consistency of generic sequential algorithms and formulate the expected minimax rate for their performance. We introduce and analyze a first algorithm called LIPO which assumes the Lipschitz constant to be known. Consistency, minimax rates for LIPO are proved, as well as fast rates under an additional Hölder like condition. An adaptive version of LIPO is also introduced for the more realistic setup where Lipschitz constant is unknown and has to be estimated along with the optimization. Similar theoretical guarantees are shown to hold for the adaptive LIPO algorithm and a numerical assessment is provided at the end of the paper to illustrate the potential of this strategy with respect to state-of-the-art methods over typical benchmark problems for global optimization.

****

On Mixed Memberships and Symmetric Nonnegative Matrix Factorizations

Xueyu Mao, Purnamrita Sarkar, Deepayan Chakrabarti

The problem of finding overlapping communities in networks has gained much attention recently. Optimization-based approaches use non-negative matrix factorization (NMF) or variants, but the global optimum cannot be provably attained in general. Model-based approaches, such as the popular mixed-membership stochastic blockmodel or MMSB (Airoldi et al., 2008), use parameters for each node to specify the overlapping communities, but standard inference techniques cannot guarantee consistency. We link the two approaches, by (a) establishing sufficient conditions for the symmetric NMF optimization to have a unique solution under MMSB, and (b) proposing a computationally efficient algorithm called GeoNMF that is provably optimal and hence consistent for a broad parameter regime. We demonstrate its accuracy on both simulated and real-world datasets.

****

Bayesian Models of Data Streams with Hierarchical Power Priors

Andrés Masegosa, Thomas D. Nielsen, Helge Langseth, Dar■■o Ramos-López, Antonio Salmerón, Anders L. Madsen

Making inferences from data streams is a pervasive problem in many modern data analysis applications. But it requires to address the problem of continuous model updating, and adapt to changes or drifts in the underlying data generating distribution. In this paper, we approach these problems from a Bayesian perspective covering general conjugate exponential models. Our proposal makes use of non-conjugate hierarchical priors to explicitly model temporal changes of the model parameters. We also derive a novel variational inference scheme which overcomes the use of non-conjugate priors while maintaining the computational efficiency of variational methods over conjugate models. The approach is validated on three real data sets over three latent variable models.

****

Just Sort It! A Simple and Effective Approach to Active Preference Learning

Lucas Maystre, Matthias Grossglauser

We address the problem of learning a ranking by using adaptively chosen pairwise comparisons. Our goal is to recover the ranking accurately but to sample the comparisons sparingly. If all comparison outcomes are consistent with the ranking, the optimal solution is to use an efficient sorting algorithm, such as Quicksort. But how do sorting algorithms behave if some comparison outcomes are inconsistent with the ranking? We give favorable guarantees for Quicksort for the popular Bradley-Terry model, under natural assumptions on the parameters. Furthermore, we empirically demonstrate that sorting algorithms lead to a very simple and effective active learning strategy: repeatedly sort the items. This strategy performs as well as state-of-the-art methods (and much better than random sampling) at a minuscule fraction of the computational cost.

****

ChoiceRank: Identifying Preferences from Node Traffic in Networks

Lucas Maystre, Matthias Grossglauser

Understanding how users navigate in a network is of high interest in many applications. We consider a setting where only aggregate node-level traffic is observed and tackle the task of learning edge transition probabilities. We cast it as a preference learning problem, and we study a model where choices follow Luce’s axiom. In this case, the $O(n)$ marginal counts of node visits are a sufficient statistic for the $O(n^2)$ transition probabilities. We show how to make the inference problem well-posed regardless of the network’s structure, and we present ChoiceRank, an iterative algorithm that scales to networks that contains billions of nodes and edges. We apply the model to two clickstream datasets and show tha t it successfully recovers the transition probabilities using only the network structure and marginal (node-level) traffic data. Finally, we also consider an application to mobility networks and apply the model to one year of rides on New York City’s bicycle-sharing system.

****

Deciding How to Decide: Dynamic Routing in Artificial Neural Networks

Mason McGill, Pietro Perona

We propose and systematically evaluate three strategies for training dynamically-routed artificial neural networks: graphs of learned transformations through which different input signals may take different paths. Though some approaches have advantages over others, the resulting networks are often qualitatively similar. We find that, in dynamically-routed networks trained to classify images, layers and branches become specialized to process distinct categories of images. Additionally, given a fixed computational budget, dynamically-routed networks tend to perform better than comparable statically-routed networks.

****

Risk Bounds for Transferring Representations With and Without Fine-Tuning

Daniel McNamara, Maria-Florina Balcan

A popular machine learning strategy is the transfer of a representation (i.e. a feature extraction function) learned on a source task to a target task. Examples include the re-use of neural network weights or word embeddings. We develop sufficient conditions for the success of this approach. If the representation learned from the source task is fixed, we identify conditions on how the tasks relate to obtain an upper bound on target task risk via a VC dimension-based argument. We then consider using the representation from the source task to construct a prior, which is fine-tuned using target task data. We give a PAC-Bayes target task risk bound in this setting under suitable conditions. We show examples of our bounds using feedforward neural networks. Our results motivate a practical approach to weight transfer, which we validate with experiments.

****

Nonnegative Matrix Factorization for Time Series Recovery From a Few Temporal Aggregates

Jiali Mei, Yohann De Castro, Yannig Goude, Georges Hébrail

Motivated by electricity consumption reconstitution, we propose a new matrix recovery method using nonnegative matrix factorization (NMF). The task tackled here is to reconstitute electricity consumption time series at a fine temporal scale from measures that are temporal aggregates of individual consumption. Contrary to existing NMF algorithms, the proposed method uses temporal aggregates as input data, instead of matrix entries. Furthermore, the proposed method is extended to take into account individual autocorrelation to provide better estimation, using a recent convex relaxation of quadratically constrained quadratic programs. Extensive experiments on synthetic and real-world electricity consumption datasets illustrate the effectiveness of the proposed method.

****

Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks

Lars Mescheder, Sebastian Nowozin, Andreas Geiger

Variational Autoencoders (VAEs) are expressive latent variable models that can be used to learn complex probability distributions from training data. However, the quality of the resulting model crucially relies on the expressiveness of the inference model. We introduce Adversarial Variational Bayes (AVB), a technique for training Variational Autoencoders with arbitrarily expressive inference models. We achieve this by introducing an auxiliary discriminative network that allows to rephrase the maximum-likelihood-problem as a two-player game, hence establishing a principled connection between VAEs and Generative Adversarial Networks (GANs). We show that in the nonparametric limit our method yields an exact maximum-likelihood assignment for the parameters of the generative model, as well as the exact posterior distribution over the latent variables given an observation. Contrary to competing approaches which combine VAEs with GANs, our approach has a clear theoretical justification, retains most advantages of standard Variation al Autoencoders and is easy to implement.

****

Efficient Orthogonal Parametrisation of Recurrent Neural Networks Using Householder Reflections
Zakaria Mhammedi, Andrew Hellicar, Ashfaqur Rahman, James Bailey

The problem of learning long-term dependencies in sequences using Recurrent Neural Networks (RNNs) is still a major challenge. Recent methods have been suggested to solve this problem by constraining the transition matrix to be unitary during training which ensures that its norm is equal to one and prevents exploding gradients. These methods either have limited expressiveness or scale poorly with the size of the network when compared with the simple RNN case, especially when using stochastic gradient descent with a small mini-batch size. Our contributions are as follows; we first show that constraining the transition matrix to be unitary is a special case of an orthogonal constraint. Then we present a new parametrisation of the transition matrix which allows efficient training of an RNN while ensuring that the matrix is always orthogonal. Our results show that the orthogonal constraint on the transition matrix applied through our parametrisation gives similar benefits to the unitary constraint, without the time complexity limitations.

****

Discovering Discrete Latent Topics with Neural Variational Inference
Yishu Miao, Edward Grefenstette, Phil Blunsom

Topic models have been widely explored as probabilistic generative models of documents. Traditional inference methods have sought closed-form derivations for updating the models, however as the expressiveness of these models grows, so does the difficulty of performing fast and accurate inference over their parameters. This paper presents alternative neural approaches to topic modelling by providing parameterisable distributions over topics which permit training by backpropagation in the framework of neural variational inference. In addition, with the help of a stick-breaking construction, we propose a recurrent network that is able to discover a notionally unbounded number of topics, analogous to Bayesian non-parametric topic models. Experimental results on the MXM Song Lyrics, 20NewsGroups and Reuters News datasets demonstrate the effectiveness and efficiency of these neural topic models.

****

Variational Boosting: Iteratively Refining Posterior Approximations
Andrew C. Miller, Nicholas J. Foti, Ryan P. Adams

We propose a black-box variational inference method to approximate intractable distributions with an increasingly rich approximating class. Our method, variational boosting, iteratively refines an existing variational approximation by solving a sequence of optimization problems, allowing a trade-off between computation time and accuracy. We expand the variational approximating class by incorporating additional covariance structure and by introducing new components to form a mixture. We apply variational boosting to synthetic and real statistical models, and show that the resulting posterior inferences compare favorably to existing variational algorithms.

****

Device Placement Optimization with Reinforcement Learning
Azalia Mirhoseini, Hieu Pham, Quoc V. Le, Benoit Steiner, Rasmus Larsen, Yuefeng Zhou, Naveen Kumar, Mohammad Norouzi, Samy Bengio, Jeff Dean

The past few years have witnessed a growth in size and computational requirements for training and inference with neural networks. Currently, a common approach to address these requirements is to use a heterogeneous distributed environment with a mixture of hardware devices such as CPUs and GPUs. Importantly, the decision of placing parts of the neural models on devices is often made by human experts based on simple heuristics and intuitions. In this paper, we propose a method which learns to optimize device placement for TensorFlow computational graphs. Key to our method is the use of a sequence-to-sequence model to predict which subsets of operations in a TensorFlow graph should run on which of the available devices. The execution time of the predicted placements is then used as the rewa rd signal to optimize the parameters of the sequence-to-sequence model. Our main result is that on Inception-V3 for ImageNet classification, and on RNN LSTM, for language modeling and neural machine translation, our model finds non-trivial device placements that outperform hand-crafted heuristics and traditional algo-rithmic methods.

****

Tight Bounds for Approximate Carathéodory and Beyond

Vahab Mirrokni, Renato Paes Leme, Adrian Vladu, Sam Chiu-wai Wong

We present a deterministic nearly-linear time algorithm for approximating any point inside a convex polytope with a sparse convex combination of the polytope’s vertices. Our result provides a constructive proof for the Approximate Carathéodory Problem, which states that any point inside a polytope contained in the ℓ_p ball of radius D can be approximated to within ϵ in ℓ_p norm by a convex combination of $O\left(D^2 p/\epsilon^2\right)$ vertices of the polytope for $p \geq 2$. While for the particular case of $p=2$, this can be achieved by the well-known Perceptron algorithm, we follow a more principled approach which generalizes to arbitrary $p\geq 2$; furthermore, this naturally extends to domains with more complicated geometry, as it is the case for providing an approximate Birkhoff-von Neumann decomposition. Secondly, we show that the sparsity bound is tight for ℓ_p norms, using an argument based on anti-concentration for the binomial distribution, thus resolving an open question posed by Barman. Experimentally, we verify that our deterministic optimization-based algorithms achieve in practice much better sparsity than previously known sampling-based algorithms. We also show how to apply our techniques to SVM training and rounding fractional points in matroid and flow polytopes.

****

Deletion-Robust Submodular Maximization: Data Summarization with “the Right to be Forgotten”

Baharan Mirzasoleiman, Amin Karbasi, Andreas Krause

How can we summarize a dynamic data stream when elements selected for the summary can be deleted at any time? This is an important challenge in online services, where the users generating the data may decide to exercise their right to restrict the service provider from using (part of) their data due to privacy concerns. Motivated by this challenge, we introduce the dynamic deletion-robust submodular maximization problem. We develop the first resilient streaming algorithm, called ROBUST-STREAMING, with a constant factor approximation guarantee to the optimum solution. We evaluate the effectiveness of our approach on several real-world applica tions, including summarizing (1) streams of geo-coordinates (2); streams of images; and (3) click-stream log data, consisting of 45 million feature vectors from a news recommendation task.

****

Prediction and Control with Temporal Segment Models

Nikhil Mishra, Pieter Abbeel, Igor Mordatch

We introduce a method for learning the dynamics of complex nonlinear systems based on deep generative models over temporal segments of states and actions. Unlike dynamics models that operate over individual discrete timesteps, we learn the distribution over future state trajectories conditioned on past state, past action, and planned future action trajectories, as well as a latent prior over action trajectories. Our approach is based on convolutional autoregressive models and variational autoencoders. It makes stable and accurate predictions over long horizons for complex, stochastic systems, effectively expressing uncertainty and modeling the effects of collisions, sensory noise, and action delays. The learned dynamics model and action prior can be used for end-to-end, fully differentiable trajectory optimization and model-based policy optimization, which we use to evaluate the performance and sample-efficiency of our method.

****

Improving Gibbs Sampler Scan Quality with DoGS

Ioannis Mitliagkas, Lester Mackey

The pairwise influence matrix of Dobrushin has long been used as an analytical tool to bound the rate of convergence of Gibbs sampling. In this work, we use Dob

rushin influence as the basis of a practical tool to certify and efficiently improve the quality of a Gibbs sampler. Our Dobrushin-optimized Gibbs samplers (DoGS) offer customized variable selection orders for a given sampling budget and variable subset of interest, explicit bounds on total variation distance to stationarity, and certifiable improvements over the standard systematic and uniform random scan Gibbs samplers. In our experiments with image segmentation, Markov chain Monte Carlo maximum likelihood estimation, and Ising model inference, DoGS consistently deliver higher-quality inferences with significantly smaller sampling budgets than standard Gibbs samplers.
****

Differentially Private Submodular Maximization: Data Summarization in Disguise
Marko Mitrovic, Mark Bun, Andreas Krause, Amin Karbasi
Many data summarization applications are captured by the general framework of submodular maximization. As a consequence, a wide range of efficient approximation algorithms have been developed. However, when such applications involve sensitive data about individuals, their privacy concerns are not automatically addressed. To remedy this problem, we propose a general and systematic study of differentially private submodular maximization. We present privacy-preserving algorithms for both monotone and non-monotone submodular maximization under cardinality, matroid, and p-extendible system constraints, with guarantees that are competitive with optimal. Along the way, we analyze a new algorithm for non-monotone submodular maximization, which is the first (even non-privately) to achieve a constant approximation ratio while running in linear time. We additionally provide two concrete experiments to validate the efficacy of these algorithms.
****

Active Learning for Top-K Rank Aggregation from Noisy Comparisons
Soheil Mohajer, Changho Suh, Adel Elmahdy
We explore an active top-K ranking problem based on pairwise comparisons that are collected possibly in a sequential manner as per our design choice. We consider two settings: (1) top-K sorting in which the goal is to recover the top-K items in order out of n items; (2) top-K partitioning where only the set of top-K items is desired. Under a fairly general model which subsumes as special cases various models (e.g., Strong Stochastic Transitivity model, BTL model and uniform noise model), we characterize upper bounds on the sample size required for top-K sorting as well as for top-K partitioning. As a consequence, we demonstrate that active ranking can offer significant multiplicative gains in sample complexity over passive ranking. Depending on the underlying stochastic noise model, such gain varies from around $\frac{\log n}{\log \log n}$ to $\frac{ n^2 \log n }{\log \log n}$. We also present an algorithm that is applicable to both settings.
****

Variational Dropout Sparsifies Deep Neural Networks
Dmitry Molchanov, Arsenii Ashukha, Dmitry Vetrov
We explore a recently proposed Variational Dropout technique that provided an elegant Bayesian interpretation to Gaussian Dropout. We extend Variational Dropout to the case when dropout rates are unbounded, propose a way to reduce the variance of the gradient estimator and report first experimental results with individual dropout rates per weight. Interestingly, it leads to extremely sparse solutions both in fully-connected and convolutional layers. This effect is similar to automatic relevance determination effect in empirical Bayes but has a number of advantages. We reduce the number of parameters up to 280 times on LeNet architectures and up to 68 times on VGG-like networks with a negligible decrease of accuracy.
****

Regularising Non-linear Models Using Feature Side-information
Amina Mollaysa, Pablo Strasser, Alexandros Kalousis
Very often features come with their own vectorial descriptions which provide detailed information about their properties. We refer to these vectorial descriptions as feature side-information. In the standard learning scenario, input is represented as a vector of features and the feature side-information is most often i

gnored or used only for feature selection prior to model fitting. We believe that feature side-information which carries information about features intrinsic property will help improve model prediction if used in a proper way during learning process. In this paper, we propose a framework that allows for the incorporation of the feature side-information during the learning of very general model families to improve the prediction performance. We control the structures of the learned models so that they reflect features’ similarities as these are defined on the basis of the side-information. We perform experiments on a number of benchmark datasets which show significant predictive performance gains, over a number of baselines, as a result of the exploitation of the side-information.

****

Coupling Distributed and Symbolic Execution for Natural Language Queries

Lili Mou, Zhengdong Lu, Hang Li, Zhi Jin

Building neural networks to query a knowledge base (a table) with natural language is an emerging research topic in deep learning. An executor for table querying typically requires multiple steps of execution because queries may have complicated structures. In previous studies, researchers have developed either fully distributed executors or symbolic executors for table querying. A distributed executor can be trained in an end-to-end fashion, but is weak in terms of execution efficiency and explicit interpretability. A symbolic executor is efficient in execution, but is very difficult to train especially at initial stages. In this paper, we propose to couple distributed and symbolic execution for natural language queries, where the symbolic executor is pretrained with the distributed executor’s intermediate execution results in a step-by-step fashion. Experiments show that our approach significantly outperforms both distributed and symbolic executors, exhibiting high accuracy, high learning efficiency, high execution efficiency, and high interpretability.

****

McGan: Mean and Covariance Feature Matching GAN

Youssef Mroueh, Tom Sercu, Vaibhava Goel

We introduce new families of Integral Probability Metrics (IPM) for training Generative Adversarial Networks (GAN). Our IPMs are based on matching statistics of distributions embedded in a finite dimensional feature space. Mean and covariance feature matching IPMs allow for stable training of GANs, which we will call McGan. McGan minimizes a meaningful loss between distributions.

****

Sequence to Better Sequence: Continuous Revision of Combinatorial Structures

Jonas Mueller, David Gifford, Tommi Jaakkola

We present a model that, after learning on observations of (sequence, outcome) pairs, can be efficiently used to revise a new sequence in order to improve its associated outcome. Our framework requires neither example improvements, nor additional evaluation of outcomes for proposed revisions. To avoid combinatorial-search over sequence elements, we specify a generative model with continuous latent factors, which is learned via joint approximate inference using a recurrent variational autoencoder (VAE) and an outcome-predicting neural network module. Under this model, gradient methods can be used to efficiently optimize the continuous latent factors with respect to inferred outcomes. By appropriately constraining this optimization and using the VAE decoder to generate a revised sequence, we ensure the revision is fundamentally similar to the original sequence, is associated with better outcomes, and looks natural. These desiderata are proven to hold with high probability under our approach, which is empirically demonstrated for revising natural language sentences.

****

Variants of RMSProp and Adagrad with Logarithmic Regret Bounds

Mahesh Chandra Mukkamala, Matthias Hein

Adaptive gradient methods have become recently very popular, in particular as they have been shown to be useful in the training of deep neural networks. In this paper we have analyzed RMSProp, originally proposed for the training of deep neural networks, in the context of online convex optimization and show $\sqrt{T}$-type regret bounds. Moreover, we propose two variants SC-Adagrad and SC-RMSProp for which we show logarithmic regret bounds for strongly convex functions. Finally, we demonstrate in the experiments that these new variants outperform other adaptive gradient techniques or stochastic gradient descent in the optimization of strongly convex functions as well as in training of deep neural networks.

****

Meta Networks

Tsendsuren Munkhdalai, Hong Yu

Neural networks have been successfully applied in applications with a large amount of labeled data. However, the task of rapid generalization on new concepts with small training data while preserving performances on previously learned ones still presents a significant challenge to neural network models. In this work, we introduce a novel meta learning method, Meta Networks (MetaNet), that learns a meta-level knowledge across tasks and shifts its inductive biases via fast parameterization for rapid generalization. When evaluated on Omniglot and Mini-ImageNet benchmarks, our MetaNet models achieve a near human-level performance and outperform the baseline approaches by up to 6\% accuracy. We demonstrate several appealing properties of MetaNet relating to generalization and continual learning.

****

Understanding the Representation and Computation of Multilayer Perceptrons: A Case Study in Speech Recognition

Tasha Nagamine, Nima Mesgarani

Despite the recent success of deep learning, the nature of the transformations they apply to the input features remains poorly understood. This study provides an empirical framework to study the encoding properties of node activations in various layers of the network, and to construct the exact function applied to each data point in the form of a linear transform. These methods are used to discern and quantify properties of feed-forward neural networks trained to map acoustic features to phoneme labels. We show a selective and nonlinear warping of the feature space, achieved by forming prototypical functions to account for the possible variation of each class. This study provides a joint framework where the properties of node activations and the functions implemented by the network can be linked together.

****

Adaptive Sampling Probabilities for Non-Smooth Optimization

Hongseok Namkoong, Aman Sinha, Steve Yadlowsky, John C. Duchi

Standard forms of coordinate and stochastic gradient methods do not adapt to structure in data; their good behavior under random sampling is predicated on uniformity in data. When gradients in certain blocks of features (for coordinate descent) or examples (for SGD) are larger than others, there is a natural structure that can be exploited for quicker convergence. Yet adaptive variants often suffer nontrivial computational overhead. We present a framework that discovers and leverages such structural properties at a low computational cost. We employ a bandit optimization procedure that “learns” probabilities for sampling coordinates or examples in (non-smooth) optimization problems, allowing us to guarantee performance close to that of the optimal stationary sampling distribution. When such structures exist, our algorithms achieve tighter convergence guarantees than their non-adaptive counterparts, and we complement our analysis with experiments on several datasets.

****

Delta Networks for Optimized Recurrent Network Computation

Daniel Neil, Jun Haeng Lee, Tobi Delbruck, Shih-Chii Liu

Many neural networks exhibit stability in their activation patterns over time in response to inputs from sensors operating under real-world conditions. By capitalizing on this property of natural signals, we propose a Recurrent Neural Network (RNN) architecture called a delta network in which each neuron transmits its value only when the change in its activation exceeds a threshold. The execution of RNNs as delta networks is attractive because their states must be stored and fetched at every timestep, unlike in convolutional neural networks (CNNs). We show that a naive run-time delta network implementation offers modest improvements on the number of memory accesses and computes, but optimized training techniques confer higher accuracy at higher speedup. With these optimizations, we demonstrate a 9X reduction in cost with negligible loss of accuracy for the TIDIGITS audio digit recognition benchmark. Similarly, on the large Wall Street Journal (WSJ) speech recognition benchmark, pretrained networks can also be greatly accelerated as delta networks and trained delta networks show a 5.7x improvement with negligible loss of accuracy. Finally, on an end-to-end CNN-RNN network trained for steering angle prediction in a driving dataset, the RNN cost can be reduced by a substantial 100X.

****

Post-Inference Prior Swapping

Willie Neiswanger, Eric Xing

While Bayesian methods are praised for their ability to incorporate useful prior knowledge, in practice, convenient priors that allow for computationally cheap or tractable inference are commonly used. In this paper, we investigate the following question: for a given model, is it possible to compute an inference result with any convenient false prior, and afterwards, given any target prior of interest, quickly transform this result into the target posterior? A potential solution is to use importance sampling (IS). However, we demonstrate that IS will fail for many choices of the target prior, depending on its parametric form and similarity to the false prior. Instead, we propose prior swapping, a method that leverages the pre-inferred false posterior to efficiently generate accurate posterior samples under arbitrary target priors. Prior swapping lets us apply less-costly inference algorithms to certain models, and incorporate new or updated prior information “post-inference”. We give theoretical guarantees about our method, and demonstrate it empirically on a number of models and priors.

****

The Loss Surface of Deep and Wide Neural Networks

Quynh Nguyen, Matthias Hein

While the optimization problem behind deep neural networks is highly non-convex, it is frequently observed in practice that training deep networks seems possible without getting stuck in suboptimal points. It has been argued that this is the case as all local minima are close to being globally optimal. We show that this is (almost) true, in fact almost all local minima are globally optimal, for a fully connected network with squared loss and analytic activation function given that the number of hidden units of one layer of the network is larger than the number of training points and the network structure from this layer on is pyramidal.

****

SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient

Lam M. Nguyen, Jie Liu, Katya Scheinberg, Martin Taká■

In this paper, we propose a StochAstic Recursive grAdient algoritHm (SARAH), as well as its practical variant SARAH+, as a novel approach to the finite-sum minimization problems. Different from the vanilla SGD and other modern stochastic methods such as SVRG, S2GD, SAG and SAGA, SARAH admits a simple recursive framework for updating stochastic gradient estimates; when comparing to SAG/SAGA, SARAH does not require a storage of past gradients. The linear convergence rate of SARAH is proven under strong convexity assumption. We also prove a linear convergence rate (in the strongly convex case) for an inner loop of SARAH, the property that SVRG does not possess. Numerical experiments demonstrate the efficiency of our algorithm.

****

Composing Tree Graphical Models with Persistent Homology Features for Clustering Mixed-Type Data

Xiuyan Ni, Novi Quadrianto, Yusu Wang, Chao Chen

Clustering data with both continuous and discrete attributes is a challenging task. Existing methods lack a principled probabilistic formulation. In this paper, we propose a clustering method based on a tree-structured graphical model to describe the generation process of mixed-type data. Our tree-structured model fact orized into a product of pairwise interactions, and thus localizes the interaction between feature variables of different types. To provide a robust clustering method based on the tree-model, we adopt a topographical view and compute peaks of the density function and their attractive basins for clustering. Furthermore, we leverage the theory from topology data analysis to adaptively merge trivial peaks into large ones in order to achieve meaningful clusterings. Our method outperforms state-of-the-art methods on mixed-type data.

****

Multichannel End-to-end Speech Recognition
Tsubasa Ochiai, Shinji Watanabe, Takaaki Hori, John R. Hershey

The field of speech recognition is in the midst of a paradigm shift: end-to-end neural networks are challenging the dominance of hidden Markov models as a core technology. Using an attention mechanism in a recurrent encoder-decoder architecture solves the dynamic time alignment problem, allowing joint end-to-end training of the acoustic and language modeling components. In this paper we extend the end-to-end framework to encompass microphone array signal processing for noise suppression and speech enhancement within the acoustic encoding network. This allows the beamforming components to be optimized jointly within the recognition architecture to improve the end-to-end speech recognition objective. Experiments on the noisy speech benchmarks (CHiME-4 and AMI) show that our multichannel end-to-end system outperformed the attention-based baseline with input from a conventional adaptive beamformer.

****

Conditional Image Synthesis with Auxiliary Classifier GANs
Augustus Odena, Christopher Olah, Jonathon Shlens

In this paper we introduce new methods for the improved training of generative adversarial networks (GANs) for image synthesis. We construct a variant of GANs employing label conditioning that results in 128×128 resolution image samples exhibiting global coherence. We expand on previous work for image quality assessment to provide two new analyses for assessing the discriminability and diversity of samples from class-conditional image synthesis models. These analyses demonstrate that high resolution samples provide class information not present in low resolution samples. Across 1000 ImageNet classes, 128×128 samples are more than twice as discriminable as artificially resized 32×32 samples. In addition, 84.7\% of the classes have samples exhibiting diversity comparable to real ImageNet data.

****

Nyström Method with Kernel K-means++ Samples as Landmarks
Dino Oglic, Thomas Gärtner

We investigate, theoretically and empirically, the effectiveness of kernel K-means++ samples as landmarks in the Nyström method for low-rank approximation of kernel matrices. Previous empirical studies (Zhang et al., 2008; Kumar et al.,2012) observe that the landmarks obtained using (kernel) K-means clustering define a good low-rank approximation of kernel matrices. However, the existing work does not provide a theoretical guarantee on the approximation error for this approach to landmark selection. We close this gap and provide the first bound on the approximation error of the Nyström method with kernel K-means++ samples as landmarks. Moreover, for the frequently used Gaussian kernel we provide a theoretically sound motivation for performing Lloyd refinements of kernel K-means++ landmarks in the instance space. We substantiate our theoretical results empirically by comparing the approach to several state-of-the-art algorithms.

****

Zero-Shot Task Generalization with Multi-Task Deep Reinforcement Learning
Junhyuk Oh, Satinder Singh, Honglak Lee, Pushmeet Kohli

As a step towards developing zero-shot task generalization capabilities in reinforcement learning (RL), we introduce a new RL problem where the agent should learn to execute sequences of instructions after learning useful skills that solve subtasks. In this problem, we consider two types of generalizations: to previously unseen instructions and to longer sequences of instructions. For generalization over unseen instructions, we propose a new objective which encourages learnin g correspondences between similar subtasks by making analogies. For generalization over sequential instructions, we present a hierarchical architecture where a meta controller learns to use the acquired skills for executing the instructions. To deal with delayed reward, we propose a new neural architecture in the meta controller that learns when to update the subtask, which makes learning more efficient. Experimental results on a stochastic 3D domain show that the proposed ideas are crucial for generalization to longer instructions as well as unseen instructions.

****

The Statistical Recurrent Unit
Junier B. Oliva, Barnabás Póczos, Jeff Schneider

Sophisticated gated recurrent neural network architectures like LSTMs and GRUs have been shown to be highly effective in a myriad of applications. We develop an un-gated unit, the statistical recurrent unit (SRU), that is able to learn long term dependencies in data by only keeping moving averages of statistics. The SRU’s architecture is simple, un-gated, and contains a comparable number of parameters to LSTMs; yet, SRUs perform favorably to more sophisticated LSTM and GRU alternatives, often outperforming one or both in various tasks. We show the efficacy of SRUs as compared to LSTMs and GRUs in an unbiased manner by optimizing respective architectures’ hyperparameters for both synthetic and real-world tasks.

****

Deep Decentralized Multi-task Multi-Agent Reinforcement Learning under Partial Observability

Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P. How, John Vian

Many real-world tasks involve multiple agents with partial observability and limited communication. Learning is challenging in these settings due to local viewpoints of agents, which perceive the world as non-stationary due to concurrently-exploring teammates. Approaches that learn specialized policies for individual tasks face problems when applied to the real world: not only do agents have to learn and store distinct policies for each task, but in practice identities of tasks are often non-observable, making these approaches inapplicable. This paper formalizes and addresses the problem of multi-task multi-agent reinforcement learning under partial observability. We introduce a decentralized single-task learning approach that is robust to concurrent interactions of teammates, and present an approach for distilling single-task policies into a unified policy that performs well across multiple related tasks, without explicit provision of task identity.

****

Algebraic Variety Models for High-Rank Matrix Completion

Greg Ongie, Rebecca Willett, Robert D. Nowak, Laura Balzano

We consider a non-linear generalization of low-rank matrix completion to the case where the data belongs to an algebraic variety, i.e., each data point is a solution to a system of polynomial equations. In this case the original matrix is possibly high-rank, but it becomes low-rank after mapping each column to a higher dimensional space of monomial features. Algebraic varieties capture a range of well-studied linear models, including affine subspaces and their union, but also quadratic and higher degree curves and surfaces. We study the sampling requirements for a general variety model with a focus on the union of affine subspaces. We propose an efficient matrix completion algorithm that minimizes a convex or non-convex surrogate of the rank of the lifted matrix. Our algorithm uses the well-known “kernel trick” to avoid working directly with the high-dimensional lifted data matrix and scales efficiently with data size. We show the proposed algorithm is able to recover synthetically generated data up to the predicted sampling complexity bounds. The algorithm also outperforms standard techniques in experiments with real data.

****

Why is Posterior Sampling Better than Optimism for Reinforcement Learning?

Ian Osband, Benjamin Van Roy

Computational results demonstrate that posterior sampling for reinforcement learning (PSRL) dramatically outperforms existing algorithms driven by optimism, suc h as UCRL2. We provide insight into the extent of this performance boost and the phenomenon that drives it. We leverage this insight to establish an $\tilde{O}(H\sqrt{SAT})$ Bayesian regret bound for PSRL in finite-horizon episodic Markov decision processes. This improves upon the best previous Bayesian regret bound of $\tilde{O}(H S \sqrt{AT})$ for any reinforcement learning algorithm. Our theoretical results are supported by extensive empirical evaluation.

****

Bidirectional Learning for Time-series Models with Hidden Units
Takayuki Osogami, Hiroshi Kajino, Taro Sekiyama

Hidden units can play essential roles in modeling time-series having long-term dependency or on-linearity but make it difficult to learn associated parameters. Here we propose a way to learn such a time-series model by training a backward model for the time-reversed time-series, where the backward model has a common set of parameters as the original (forward) model. Our key observation is that only a subset of the parameters is hard to learn, and that subset is complementary between the forward model and the backward model. By training both of the two models, we can effectively learn the values of the parameters that are hard to learn if only either of the two models is trained. We apply bidirectional learning to a dynamic Boltzmann machine extended with hidden units. Numerical experiments with synthetic and real datasets clearly demonstrate advantages of bidirectional learning.

****

Count-Based Exploration with Neural Density Models
Georg Ostrovski, Marc G. Bellemare, Aäron Oord, Rémi Munos

Bellemare et al. (2016) introduced the notion of a pseudo-count, derived from a density model, to generalize count-based exploration to non-tabular reinforcement learning. This pseudo-count was used to generate an exploration bonus for a DQN agent and combined with a mixed Monte Carlo update was sufficient to achieve state of the art on the Atari 2600 game Montezuma’s Revenge. We consider two questions left open by their work: First, how important is the quality of the density model for exploration? Second, what role does the Monte Carlo update play in exploration? We answer the first question by demonstrating the use of PixelCNN, an advanced neural density model for images, to supply a pseudo-count. In particular, we examine the intrinsic difficulties in adapting Bellemare et al.’s approach when assumptions about the model are violated. The result is a more practical and general algorithm requiring no special apparatus. We combine PixelCNN pseudo-counts with different agent architectures to dramatically improve the state of the art on several hard Atari games. One surprising finding is that the mixed Monte Carlo update is a powerful facilitator of exploration in the sparsest of settings, including Montezuma’s Revenge.

****

Dictionary Learning Based on Sparse Distribution Tomography
Pedram Pad, Farnood Salehi, Elisa Celis, Patrick Thiran, Michael Unser

We propose a new statistical dictionary learning algorithm for sparse signals that is based on an α-stable innovation model. The parameters of the underlying model—that is, the atoms of the dictionary, the sparsity index α and the dispersion of the transform-domain coefficients—are recovered using a new type of probability distribution tomography. Specifically, we drive our estimator with a series of random projections of the data, which results in an efficient algorithm. Moreover, since the projections are achieved using linear combinations, we can invoke the generalized central limit theorem to justify the use of our method for sparse signals that are not necessarily α-stable. We evaluate our algorithm by performing two types of experiments: image in-painting and image denoising. In both cases, we find that our approach is competitive with state-of-the-art dictionary learning techniques. Beyond the algorithm itself, two aspects of this study are interesting in their own right. The first is our statistical formulation of the problem, which unifies the topics of dictionary learning and independent component analysis. The second is a generalization of a classical theorem about isometries of ℓ_p-norms that constitutes the foundation of our approach.

****

Stochastic Bouncy Particle Sampler

Ari Pakman, Dar Gilboa, David Carlson, Liam Paninski

We introduce a stochastic version of the non-reversible, rejection-free Bouncy Particle Sampler (BPS), a Markov process whose sample trajectories are piecewise linear, to efficiently sample Bayesian posteriors in big datasets. We prove that in the BPS no bias is introduced by noisy evaluations of the log-likelihood gradient. On the other hand, we argue that efficiency considerations favor a small, controllable bias, in exchange for faster mixing. We introduce a simple method that controls this trade-off. We illustrate these ideas in several examples which outperform previous approaches.

****

A Birth-Death Process for Feature Allocation

Konstantina Palla, David Knowles, Zoubin Ghahramani

We propose a Bayesian nonparametric prior over feature allocations for sequential data, the birth-death feature allocation process (BDFP). The BDFP models the evolution of the feature allocation of a set of N objects across a covariate (e.g.~time) by creating and deleting features. A BDFP is exchangeable, projective, stationary and reversible, and its equilibrium distribution is given by the Indian buffet process (IBP). We show that the Beta process on an extended space is the de Finetti mixing distribution underlying the BDFP. Finally, we present the finite approximation of the BDFP, the Beta Event Process (BEP), that permits simplified inference. The utility of the BDFP as a prior is demonstrated on real world dynamic genomics and social network data.

****

Prediction under Uncertainty in Sparse Spectrum Gaussian Processes with Applications to Filtering and Control

Yunpeng Pan, Xinyan Yan, Evangelos A. Theodorou, Byron Boots

Sparse Spectrum Gaussian Processes (SSGPs) are a powerful tool for scaling Gaussian processes (GPs) to large datasets. Existing SSGP algorithms for regression assume deterministic inputs, precluding their use in many real-world robotics and engineering applications where accounting for input uncertainty is crucial. We address this problem by proposing two analytic moment-based approaches with closed-form expressions for SSGP regression with uncertain inputs. Our methods are more general and scalable than their standard GP counterparts, and are naturally applicable to multi-step prediction or uncertainty propagation. We show that efficient algorithms for Bayesian filtering and stochastic model predictive control can use these methods, and we evaluate our algorithms with comparative analyses and both real-world and simulated experiments.

****

Clustering by Sum of Norms: Stochastic Incremental Algorithm, Convergence and Cluster Recovery

Ashkan Panahi, Devdatt Dubhashi, Fredrik D. Johansson, Chiranjib Bhattacharyya

Standard clustering methods such as K-means, Gaussian mixture models, and hierarchical clustering are beset by local minima, which are sometimes drastically suboptimal. Moreover the number of clusters K must be known in advance. The recently introduced the sum-of-norms (SON) or Clusterpath convex relaxation of k-means and hierarchical clustering shrinks cluster centroids toward one another and ensure a unique global minimizer. We give a scalable stochastic incremental algorithm based on proximal iterations to solve the SON problem with convergence guarantees. We also show that the algorithm recovers clusters under quite general conditions which have a similar form to the unifying proximity condition introduced in the approximation algorithms community (that covers paradigm cases such as Gaussian mixtures and planted partition models). We give experimental results to confirm that our algorithm scales much better than previous methods while producing clusters of comparable quality.

****

Curiosity-driven Exploration by Self-supervised Prediction

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, Trevor Darrell

In many real-world scenarios, rewards extrinsic to the agent are extremely spars e, or absent altogether. In such cases, curiosity can serve as an intrinsic reward signal to enable the agent to explore its environment and learn skills that might be useful later in its life. We formulate curiosity as the error in an agent’s ability to predict the consequence of its own actions in a visual feature space learned by a self-supervised inverse dynamics model. Our formulation scales to high-dimensional continuous state spaces like images, bypasses the difficulties of directly predicting pixels, and, critically, ignores the aspects of the environment that cannot affect the agent. The proposed approach is evaluated in two environments: VizDoom and Super Mario Bros. Three broad settings are investigated: 1) sparse extrinsic reward, where curiosity allows for far fewer interactions with the environment to reach the goal; 2) exploration with no extrinsic reward, where curiosity pushes the agent to explore more efficiently; and 3) generalization to unseen scenarios (e.g. new levels of the same game) where the knowledge gained from earlier experience helps the agent explore new places much faster than starting from scratch.

****

Asynchronous Distributed Variational Gaussian Process for Regression

Hao Peng, Shandian Zhe, Xiao Zhang, Yuan Qi

Gaussian processes (GPs) are powerful non-parametric function estimators. However, their applications are largely limited by the expensive computational cost of the inference procedures. Existing stochastic or distributed synchronous variational inferences, although have alleviated this issue by scaling up GPs to millions of samples, are still far from satisfactory for real-world large applications, where the data sizes are often orders of magnitudes larger, say, billions. To solve this problem, we propose ADVGP, the first Asynchronous Distributed Variational Gaussian Process inference for regression, on the recent large-scale machine learning platform, PARAMETER SERVER. ADVGP uses a novel, flexible variational framework based on a weight space augmentation, and implements the highly efficient, asynchronous proximal gradient optimization. While maintaining comparable or better predictive performance, ADVGP greatly improves upon the efficiency of the existing variational methods. With ADVGP, we effortlessly scale up GP regression to a real-world application with billions of samples and demonstrate an excellent, superior prediction accuracy to the popular linear models.

****

Geometry of Neural Network Loss Surfaces via Random Matrix Theory

Jeffrey Pennington, Yasaman Bahri

Understanding the geometry of neural network loss surfaces is important for the development of improved optimization algorithms and for building a theoretical understanding of why deep learning works. In this paper, we study the geometry in terms of the distribution of eigenvalues of the Hessian matrix at critical points of varying energy. We introduce an analytical framework and a set of tools from random matrix theory that allow us to compute an approximation of this distribution under a set of simplifying assumptions. The shape of the spectrum depends strongly on the energy and another key parameter, ϕ, which measures the ratio of parameters to data points. Our analysis predicts and numerical simulations support that for critical points of small index, the number of negative eigenvalues scales like the 3/2 power of the energy. We leave as an open problem an explanation for our observation that, in the context of a certain memorization task, the energy of minimizers is well-approximated by the function $1/2(1-\phi)^2$.

****

Multi-task Learning with Labeled and Unlabeled Tasks

Anastasia Pentina, Christoph H. Lampert

In multi-task learning, a learner is given a collection of prediction tasks and needs to solve all of them. In contrast to previous work, which required that annotated training data must be available for all tasks, we consider a new setting, in which for some tasks, potentially most of them, only unlabeled training data is provided. Consequently, to solve all tasks, information must be transferred between tasks with labels and tasks without labels. Focusing on an instance-based transfer method we analyze two variants of this setting: when the set of labe led tasks is fixed, and when it can be actively selected by the learner. We state and prove a generalization bound that covers both scenarios and derive from it an algorithm for making the choice of labeled tasks (in the active case) and for transferring information between the tasks in a principled way. We also illustrate the effectiveness of the algorithm on synthetic and real data.

****

Robust Adversarial Reinforcement Learning

Lerrel Pinto, James Davidson, Rahul Sukthankar, Abhinav Gupta

Deep neural networks coupled with fast simulation and improved computational speeds have led to recent successes in the field of reinforcement learning (RL). However, most current RL-based approaches fail to generalize since: (a) the gap between simulation and real world is so large that policy-learning approaches fail to transfer; (b) even if policy learning is done in real world, the data scarcity leads to failed generalization from training to test scenarios (e.g., due to different friction or object masses). Inspired from H-infinity control methods, we note that both modeling errors and differences in training and test scenarios can just be viewed as extra forces/disturbances in the system. This paper proposes the idea of robust adversarial reinforcement learning (RARL), where we train an agent to operate in the presence of a destabilizing adversary that applies disturbance forces to the system. The jointly trained adversary is reinforced – that is, it learns an optimal destabilization policy. We formulate the policy learning as a zero-sum, minimax objective function. Extensive experiments in multiple environments (InvertedPendulum, HalfCheetah, Swimmer, Hopper, Walker2d and Ant) conclusively demonstrate that our method (a) improves training stability; (b) is robust to differences in training/test conditions; and c) outperform the baseline even in the absence of the adversary.

****

Neural Episodic Control

Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adrià Puigdomènech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, Charles Blundell

Deep reinforcement learning methods attain super-human performance in a wide range of environments. Such methods are grossly inefficient, often taking orders of magnitudes more data than humans to achieve reasonable performance. We propose Neural Episodic Control: a deep reinforcement learning agent that is able to rapidly assimilate new experiences and act upon them. Our agent uses a semi-tabular representation of the value function: a buffer of past experience containing slowly changing state representations and rapidly updated estimates of the value function. We show across a wide range of environments that our agent learns significantly faster than other state-of-the-art, general purpose deep reinforcement learning agents.

****

Online and Linear-Time Attention by Enforcing Monotonic Alignments

Colin Raffel, Minh-Thang Luong, Peter J. Liu, Ron J. Weiss, Douglas Eck

Recurrent neural network models with an attention mechanism have proven to be extremely effective on a wide variety of sequence-to-sequence problems. However, the fact that soft attention mechanisms perform a pass over the entire input sequence when producing each element in the output sequence precludes their use in online settings and results in a quadratic time complexity. Based on the insight that the alignment between input and output sequence elements is monotonic in many problems of interest, we propose an end-to-end differentiable method for learning monotonic alignments which, at test time, enables computing attention online and in linear time. We validate our approach on sentence summarization, machine translation, and online speech recognition problems and achieve results competitive with existing sequence-to-sequence models.

****

On the Expressive Power of Deep Neural Networks

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, Jascha Sohl-Dickstein

We propose a new approach to the problem of neural network expressivity, which seeks to characterize how structural properties of a neural network family affect the functions it is able to compute. Our approach is based on an interrelated s et of measures of expressivity, unified by the novel notion of trajectory length, which measures how the output of a network changes as the input sweeps along a one-dimensional path. Our findings show that: (1) The complexity of the computed function grows exponentially with depth (2) All weights are not equal: trained networks are more sensitive to their lower (initial) layer weights (3) Trajectory regularization is a simpler alternative to batch normalization, with the same performance.

****

Estimating the unseen from multiple populations

Aditi Raghunathan, Gregory Valiant, James Zou

Given samples from a distribution, how many new elements should we expect to find if we keep on sampling this distribution? This is an important and actively studied problem, with many applications ranging from species estimation to genomics. We generalize this extrapolation and related unseen estimation problems to the multiple population setting, where population j has an unknown distribution D_j from which we observe n_j samples. We derive an optimal estimator for the total number of elements we expect to find among new samples across the populations. Surprisingly, we prove that our estimator’s accuracy is independent of the number of populations. We also develop an efficient optimization algorithm to solve the more general problem of estimating multi-population frequency distributions. We validate our methods and theory through extensive experiments. Finally, on a real dataset of human genomes across multiple ancestries, we demonstrate how our approach for unseen estimation can enable cohort designs that can discover interesting mutations with greater efficiency.

****

Coherence Pursuit: Fast, Simple, and Robust Subspace Recovery

Mostafa Rahmani, George Atia

This paper presents a remarkably simple, yet powerful, algorithm for robust Principal Component Analysis (PCA). In the proposed approach, an outlier is set apart from an inlier by comparing their coherence with the rest of the data points. As inliers lie on a low dimensional subspace, they are likely to have strong mutual coherence provided there are enough inliers. By contrast, outliers do not typically admit low dimensional structures, wherefore an outlier is unlikely to bear strong resemblance with a large number of data points. The mutual coherences are computed by forming the Gram matrix of normalized data points. Subsequently, the subspace is recovered from the span of a small subset of the data points that exhibit strong coherence with the rest of the data. As coherence pursuit only involves one simple matrix multiplication, it is significantly faster than the state of-the-art robust PCA algorithms. We provide a mathematical analysis of the proposed algorithm under a random model for the distribution of the inliers and outliers. It is shown that the proposed method can recover the correct subspace even if the data is predominantly outliers. To the best of our knowledge, this is the first provable robust PCA algorithm that is simultaneously non-iterative, can tolerate a large number of outliers and is robust to linearly dependent outliers.

****

Innovation Pursuit: A New Approach to the Subspace Clustering Problem

Mostafa Rahmani, George Atia

This paper presents a new scalable approach, termed Innovation Pursuit (iPursuit), to the problem of subspace clustering. iPursuit rests on a new geometrical idea whereby each subspace is identified based on its novelty with respect to the other subspaces. The subspaces are identified consecutively by solving a series of simple linear optimization problems, each searching for a direction of innovation in the span of the data. A detailed mathematical analysis is provided establishing sufficient conditions for the proposed approach to correctly cluster the data points. Moreover, the proposed direction search approach can be integrated with spectral clustering to yield a new variant of spectral-clustering-based algorithms. Remarkably, the proposed approach can provably yield exact clustering even when the subspaces have significant intersections. The numerical simulations demonstrate that iPursuit can often outperform the state-of-the-art subspace c lustering algorithms – more so for subspaces with significant intersections – along with substantial reductions in computational complexity.

****

High Dimensional Bayesian Optimization with Elastic Gaussian Process

Santu Rana, Cheng Li, Sunil Gupta, Vu Nguyen, Svetha Venkatesh

Bayesian optimization is an efficient way to optimize expensive black-box functions such as designing a new product with highest quality or hyperparameter tuning of a machine learning algorithm. However, it has a serious limitation when the parameter space is high-dimensional as Bayesian optimization crucially depends on solving a global optimization of a surrogate utility function in the same sized dimensions. The surrogate utility function, known commonly as acquisition function is a continuous function but can be extremely sharp at high dimension - having only a few peaks marooned in a large terrain of almost flat surface. Global optimization algorithms such as DIRECT are infeasible at higher dimensions and gradient-dependent methods cannot move if initialized in the flat terrain. We propose an algorithm that enables local gradient-dependent algorithms to move through the flat terrain by using a sequence of gross-to-finer Gaussian process priors on the objective function as we leverage two underlying facts - a) there exists a large enough length-scales for which the acquisition function can be made to have a significant gradient at any location in the parameter space, and b) the extrema of the consecutive acquisition functions are close although they are different only due to a small difference in the length-scales. Theoretical guarantees are provided and experiments clearly demonstrate the utility of the proposed method at high dimension using both benchmark test functions and real-world case studies.

****

Equivariance Through Parameter-Sharing

Siamak Ravanbakhsh, Jeff Schneider, Barnabás Póczos

We propose to study equivariance in deep neural networks through parameter symmetries. In particular, given a group G that acts discretely on the input and output of a standard neural network layer, we show that its equivariance is linked to the symmetry group of network parameters. We then propose two parameter-sharing scheme to induce the desirable symmetry on the parameters of the neural network. Under some conditions on the action of G, our procedure for tying the parameters achieves G-equivariance and guarantees sensitivity to all other permutation groups outside of G.

****

Large-Scale Evolution of Image Classifiers

Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc V. Le, Alexey Kurakin

Neural networks have proven effective at solving difficult problems but designing their architectures can be challenging, even for image classification problems alone. Our goal is to minimize human participation, so we employ evolutionary algorithms to discover such networks automatically. Despite significant computational requirements, we show that it is now possible to evolve models with accuracies within the range of those published in the last year. Specifically, we employ simple evolutionary techniques at unprecedented scales to discover models for the CIFAR-10 and CIFAR-100 datasets, starting from trivial initial conditions and reaching accuracies of 94.6\% (95.6\% for ensemble) and 77.0\%, respectively. To do this, we use novel and intuitive mutation operators that navigate large search spaces; we stress that no human participation is required once evolution starts and that the output is a fully-trained model. Throughout this work, we place special emphasis on the repeatability of results, the variability in the outcomes and the computational requirements.

****

Parallel Multiscale Autoregressive Density Estimation

Scott Reed, Aäron Oord, Nal Kalchbrenner, Sergio Gómez Colmenarejo, Ziyu Wang, Yutian Chen, Dan Belov, Nando Freitas

PixelCNN achieves state-of-the-art results in density estimation for natural images. Although training is fast, inference is costly, requiring one network evalu ation per pixel; O(N) for N pixels. This can be sped up by caching activations, but still involves generating each pixel sequentially. In this work, we propose a parallelized PixelCNN that allows more efficient inference by modeling certain pixel groups as conditionally independent. Our new PixelCNN model achieves competitive density estimation and orders of magnitude speedup – O(log N) sampling instead of O(N) – enabling the practical generation of 512x512 images. We evaluate the model on class-conditional image generation, text-to-image synthesis, and action-conditional video generation, showing that our model achieves the best results among non-pixel-autoregressive density models that allow efficient sampling.

****

Real-Time Adaptive Image Compression

Oren Rippel, Lubomir Bourdev

We present a machine learning-based approach to lossy image compression which outperforms all existing codecs, while running in real-time. Our algorithm typically produces file sizes 3 times smaller than JPEG, 2.5 times smaller than JPEG 2000, and 2.3 times smaller than WebP on datasets of generic images across a spectrum of quality levels. At the same time, our codec is designed to be lightweight and deployable: for example, it can encode or decode the Kodak dataset in less than 10ms per image on GPU. Our architecture is an autoencoder featuring pyramidal analysis, an adaptive coding module, and regularization of the expected codelength. We also supplement our approach with adversarial training specialized towards use in a compression setting: this enables us to produce visually pleasing reconstructions for very low bitrates.

****

Active Learning for Accurate Estimation of Linear Models

Carlos Riquelme, Mohammad Ghavamzadeh, Alessandro Lazaric

We explore the sequential decision making problem where the goal is to estimate uniformly well a number of linear models, given a shared budget of random contexts independently sampled from a known distribution. The decision maker must query one of the linear models for each incoming context, and receives an observation corrupted by noise levels that are unknown, and depend on the model instance. We present Trace-UCB, an adaptive allocation algorithm that learns the noise levels while balancing contexts accordingly across the different linear functions, and derive guarantees for simple regret in both expectation and high-probability. Finally, we extend the algorithm and its guarantees to high dimensional settings, where the number of linear models times the dimension of the contextual space is higher than the total budget of samples. Simulations with real data suggest that Trace-UCB is remarkably robust, outperforming a number of baselines even when its assumptions are violated.

****

Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study

Samuel Ritter, David G. T. Barrett, Adam Santoro, Matt M. Botvinick

Deep neural networks (DNNs) have advanced performance on a wide range of complex tasks, rapidly outpacing our understanding of the nature of their solutions. While past work sought to advance our understanding of these models, none has made use of the rich history of problem descriptions, theories, and experimental methods developed by cognitive psychologists to study the human mind. To explore the potential value of these tools, we chose a well-established analysis from developmental psychology that explains how children learn word labels for objects, and applied that analysis to DNNs. Using datasets of stimuli inspired by the original cognitive psychology experiments, we find that state-of-the-art one shot learning models trained on ImageNet exhibit a similar bias to that observed in humans: they prefer to categorize objects according to shape rather than color. The magnitude of this shape bias varies greatly among architecturally identical, but differently seeded models, and even fluctuates within seeds throughout training, despite nearly equivalent classification performance. These results demonstrate the capability of tools from cognitive psychology for exposing hidden computational properties of DNNs, while concurrently providing us with a computational model for human word learning.

****

Pain-Free Random Differential Privacy with Sensitivity Sampling
Benjamin I. P. Rubinstein, Francesco Aldà
Popular approaches to differential privacy, such as the Laplace and exponential mechanisms, calibrate randomised smoothing through global sensitivity of the target non-private function. Bounding such sensitivity is often a prohibitively complex analytic calculation. As an alternative, we propose a straightforward sampler for estimating sensitivity of non-private mechanisms. Since our sensitivity estimates hold with high probability, any mechanism that would be (ϵ,δ)-differentially private under bounded global sensitivity automatically achieves (ϵ,δ,γ)-random differential privacy (Hall et al. 2012), without any target-specific calculations required. We demonstrate on worked example learners how our usable approach adopts a naturally-relaxed privacy guarantee, while achieving more accurate releases even for non-private functions that are black-box computer programs.

****

Enumerating Distinct Decision Trees
Salvatore Ruggieri
The search space for the feature selection problem in decision tree learning is the lattice of subsets of the available features. We provide an exact enumeration procedure of the subsets that lead to all and only the distinct decision trees. The procedure can be adopted to prune the search space of complete and heuristics search methods in wrapper models for feature selection. Based on this, we design a computational optimization of the sequential backward elimination heuristics with a performance improvement of up to 100X.

****

Bayesian Boolean Matrix Factorisation
Tammo Rukat, Chris C. Holmes, Michalis K. Titsias, Christopher Yau
Boolean matrix factorisation aims to decompose a binary data matrix into an approximate Boolean product of two low rank, binary matrices: one containing meaningful patterns, the other quantifying how the observations can be expressed as a combination of these patterns. We introduce the OrMachine, a probabilistic generative model for Boolean matrix factorisation and derive a Metropolised Gibbs sampler that facilitates efficient parallel posterior inference. On real world and simulated data, our method outperforms all currently existing approaches for Boolean matrix factorisation and completion. This is the first method to provide full posterior inference for Boolean Matrix factorisation which is relevant in applications, e.g. for controlling false positive rates in collaborative filtering and, crucially, improves the interpretability of the inferred patterns. The proposed algorithm scales to large datasets as we demonstrate by analysing single cell gene expression data in 1.3 million mouse brain cells across 11 thousand genes on commodity hardware.

****

Depth-Width Tradeoffs in Approximating Natural Functions with Neural Networks
Itay Safran, Ohad Shamir
We provide several new depth-based separation results for feed-forward neural networks, proving that various types of simple and natural functions can be better approximated using deeper networks than shallower ones, even if the shallower networks are much larger. This includes indicators of balls and ellipses; non-linear functions which are radial with respect to the L_1 norm; and smooth non-linear functions. We also show that these gaps can be observed experimentally: Increasing the depth indeed allows better learning than increasing width, when training neural networks to learn an indicator of a unit ball.

****

Asymmetric Tri-training for Unsupervised Domain Adaptation
Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada
It is important to apply models trained on a large number of labeled samples to different domains because collecting many labeled samples in various domains is expensive. To learn discriminative representations for the target domain, we assume that artificially labeling the target samples can result in a good represent

ation. Tri-training leverages three classifiers equally to provide pseudo-labels to unlabeled samples; however, the method does not assume labeling samples generated from a different domain. In this paper, we propose the use of an asymmetric tri-training method for unsupervised domain adaptation, where we assign pseudo-labels to unlabeled samples and train the neural networks as if they are true labels. In our work, we use three networks asymmetrically, and by asymmetric, we mean that two networks are used to label unlabeled target samples, and one network is trained by the pseudo-labeled samples to obtain target-discriminative representations. Our proposed method was shown to achieve a state-of-the-art performance on the benchmark digit recognition datasets for domain adaptation.

****

Semi-Supervised Classification Based on Classification from Positive and Unlabeled Data

Tomoya Sakai, Marthinus Christoffel Plessis, Gang Niu, Masashi Sugiyama

Most of the semi-supervised classification methods developed so far use unlabeled data for regularization purposes under particular distributional assumptions such as the cluster assumption. In contrast, recently developed methods of classification from positive and unlabeled data (PU classification) use unlabeled data for risk evaluation, i.e., label information is directly extracted from unlabeled data. In this paper, we extend PU classification to also incorporate negative data and propose a novel semi-supervised learning approach. We establish generalization error bounds for our novel methods and show that the bounds decrease with respect to the number of unlabeled data without the distributional assumptions that are required in existing semi-supervised learning methods. Through experiments, we demonstrate the usefulness of the proposed methods.

****

Analytical Guarantees on Numerical Precision of Deep Neural Networks

Charbel Sakr, Yongjune Kim, Naresh Shanbhag

The acclaimed successes of neural networks often overshadow their tremendous complexity. We focus on numerical precision – a key parameter defining the complexity of neural networks. First, we present theoretical bounds on the accuracy in presence of limited precision. Interestingly, these bounds can be computed via the back-propagation algorithm. Hence, by combining our theoretical analysis and the back-propagation algorithm, we are able to readily determine the minimum precision needed to preserve accuracy without having to resort to time-consuming fixed-point simulations. We provide numerical evidence showing how our approach allows us to maintain high accuracy but with lower complexity than state-of-the-art binary networks.

****

Hierarchy Through Composition with Multitask LMDPs

Andrew M. Saxe, Adam C. Earle, Benjamin Rosman

Hierarchical architectures are critical to the scalability of reinforcement learning methods. Most current hierarchical frameworks execute actions serially, with macro-actions comprising sequences of primitive actions. We propose a novel alternative to these control hierarchies based on concurrent execution of many actions in parallel. Our scheme exploits the guaranteed concurrent compositionality provided by the linearly solvable Markov decision process (LMDP) framework, which naturally enables a learning agent to draw on several macro-actions simultaneously to solve new tasks. We introduce the Multitask LMDP module, which maintains a parallel distributed representation of tasks and may be stacked to form deep hierarchies abstracted in space and time.

****

Optimal Algorithms for Smooth and Strongly Convex Distributed Optimization in Networks

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, Laurent Massoulié

In this paper, we determine the optimal convergence rates for strongly convex and smooth distributed optimization in two settings: centralized and decentralized communications over a network. For centralized (i.e. master/slave) algorithms, we show that distributing Nesterov’s accelerated gradient descent is optimal and achieves a precision $\varepsilon > 0$ in time $O(\sqrt{\kappa_g}(1+\Delta\tau)

\ln(1/\varepsilon))$, where κ_g is the condition number of the (global) function to optimize, Δ is the diameter of the network, and τ (resp. 1) is the time needed to communicate values between two neighbors (resp. perform local computations). For decentralized algorithms based on gossip, we provide the first optimal algorithm, called the multi-step dual accelerated (MSDA) method, that achieves a precision $\varepsilon > 0$ in time $O(\sqrt{\kappa_l}(1+\frac{\tau}{\sqrt{\gamma}})\ln(1/\varepsilon))$, where κ_l is the condition number of the local functions and γ is the (normalized) eigengap of the gossip matrix used for communication between nodes. We then verify the efficiency of MSDA against state-of-the-art methods for two problems: least-squares regression and classification by logistic regression.

****

Adapting Kernel Representations Online Using Submodular Maximization
Matthew Schlegel, Yangchen Pan, Jiecao Chen, Martha White
Kernel representations provide a nonlinear representation, through similarities to prototypes, but require only simple linear learning algorithms given those prototypes. In a continual learning setting, with a constant stream of observations, it is critical to have an efficient mechanism for sub-selecting prototypes amongst observations. In this work, we develop an approximately submodular criterion for this setting, and an efficient online greedy submodular maximization algorithm for optimizing the criterion. We extend streaming submodular maximization algorithms to continual learning, by removing the need for multiple passes—which is infeasible—and instead introducing the idea of coverage time. We propose a general block-diagonal approximation for the greedy update with our criterion, that enables updates linear in the number of prototypes. We empirically demonstrate the effectiveness of this approximation, in terms of approximation quality, significant runtime improvements, and effective prediction performance.

****

Developing Bug-Free Machine Learning Systems With Formal Mathematics
Daniel Selsam, Percy Liang, David L. Dill
Noisy data, non-convex objectives, model misspecification, and numerical instability can all cause undesired behaviors in machine learning systems. As a result, detecting actual implementation errors can be extremely difficult. We demonstrate a methodology in which developers use an interactive proof assistant to both implement their system and to state a formal theorem defining what it means for their system to be correct. The process of proving this theorem interactively in the proof assistant exposes all implementation errors since any error in the program would cause the proof to fail. As a case study, we implement a new system, Certigrad, for optimizing over stochastic computation graphs, and we generate a formal (i.e. machine-checkable) proof that the gradients sampled by the system are unbiased estimates of the true mathematical gradients. We train a variational autoencoder using Certigrad and find the performance comparable to training the same model in TensorFlow.

****

Identifying Best Interventions through Online Importance Sampling
Rajat Sen, Karthikeyan Shanmugam, Alexandros G. Dimakis, Sanjay Shakkottai
Motivated by applications in computational advertising and systems biology, we consider the problem of identifying the best out of several possible soft interventions at a source node V in an acyclic causal directed graph, to maximize the expected value of a target node Y (located downstream of V). Our setting imposes a fixed total budget for sampling under various interventions, along with cost constraints on different types of interventions. We pose this as a best arm identification bandit problem with K arms, where each arm is a soft intervention at V and leverage the information leakage among the arms to provide the first gap dependent error and simple regret bounds for this problem. Our results are a significant improvement over the traditional best arm identification results. We empirically show that our algorithms outperform the state of the art in the Flow Cytometry data-set, and also apply our algorithm for model interpretation of the Inception-v3 deep net that classifies images.

****

Failures of Gradient-Based Deep Learning

Shai Shalev-Shwartz, Ohad Shamir, Shaked Shammah

In recent years, Deep Learning has become the go-to solution for a broad range of applications, often outperforming state-of-the-art. However, it is important, for both theoreticians and practitioners, to gain a deeper understanding of the difficulties and limitations associated with common approaches and algorithms. We describe four types of simple problems, for which the gradient-based algorithms commonly used in deep learning either fail or suffer from significant difficulties. We illustrate the failures through practical experiments, and provide theoretical insights explaining their source, and how they might be remedied.

****

Estimating individual treatment effect: generalization bounds and algorithms

Uri Shalit, Fredrik D. Johansson, David Sontag

There is intense interest in applying machine learning to problems of causal inference in fields such as healthcare, economics and education. In particular, individual-level causal inference has important applications such as precision medicine. We give a new theoretical analysis and family of algorithms for predicting individual treatment effect (ITE) from observational data, under the assumption known as strong ignorability. The algorithms learn a “balanced” representation such that the induced treated and control distributions look similar, and we give a novel and intuitive generalization-error bound showing the expected ITE estimation error of a representation is bounded by a sum of the standard generalization-error of that representation and the distance between the treated and control distributions induced by the representation. We use Integral Probability Metrics to measure distances between distributions, deriving explicit bounds for the Wasserstein and Maximum Mean Discrepancy (MMD) distances. Experiments on real and simulated data show the new algorithms match or outperform the state-of-the-art.

****

Online Learning with Local Permutations and Delayed Feedback

Ohad Shamir, Liran Szlak

We propose an Online Learning with Local Permutations (OLLP) setting, in which the learner is allowed to slightly permute the order of the loss functions generated by an adversary. On one hand, this models natural situations where the exact order of the learner’s responses is not crucial, and on the other hand, might allow better learning and regret performance, by mitigating highly adversarial loss sequences. Also, with random permutations, this can be seen as a setting interpolating between adversarial and stochastic losses. In this paper, we consider the applicability of this setting to convex online learning with delayed feedback, in which the feedback on the prediction made in round t arrives with some delay τ. With such delayed feedback, the best possible regret bound is well-known to be $O(\sqrt{\tau T})$. We prove that by being able to permute losses by a distance of at most M (for $M\geq \tau$), the regret can be improved to $O(\sqrt{T}(1+\sqrt{\tau^2/M}))$, using a Mirror-Descent based algorithm which can be applied for both Euclidean and non-Euclidean geometries. We also prove a lower bound, showing that for $M<\tau/3$, it is impossible to improve the standard $O(\sqrt{\tau T})$ regret bound by more than constant factors. Finally, we provide some experiments validating the performance of our algorithm.

****

Orthogonalized ALS: A Theoretically Principled Tensor Decomposition Algorithm for Practical Use

Vatsal Sharan, Gregory Valiant

The popular Alternating Least Squares (ALS) algorithm for tensor decomposition is efficient and easy to implement, but often converges to poor local optima—particularly when the weights of the factors are non-uniform. We propose a modification of the ALS approach that is as efficient as standard ALS, but provably recovers the true factors with random initialization under standard incoherence assumptions on the factors of the tensor. We demonstrate the significant practical superiority of our approach over traditional ALS for a variety of tasks on synthetic data—including tensor factorization on exact, noisy and over-complete tensors

, as well as tensor completion—and for computing word embeddings from a third-order word tri-occurrence tensor.

****

Differentially Private Ordinary Least Squares

Or Sheffet

Linear regression is one of the most prevalent techniques in machine learning; however, it is also common to use linear regression for its explanatory capabilities rather than label prediction. Ordinary Least Squares (OLS) is often used in statistics to establish a correlation between an attribute (e.g. gender) and a label (e.g. income) in the presence of other (potentially correlated) features. OLS assumes a particular model that randomly generates the data, and derives t-values — representing the likelihood of each real value to be the true correlation. Using t-values, OLS can release a confidence interval, which is an interval on the reals that is likely to contain the true correlation; and when this interval does not intersect the origin, we can reject the null hypothesis as it is likely that the true correlation is non-zero. Our work aims at achieving similar guarantees on data under differentially private estimators. First, we show that for well-spread data, the Gaussian Johnson-Lindenstrauss Transform (JLT) gives a very good approximation of t-values; secondly, when JLT approximates Ridge regression (linear regression with l_2-regularization) we derive, under certain conditions, confidence intervals using the projected data; lastly, we derive, under different conditions, confidence intervals for the “Analyze Gauss” algorithm (Dwork et al 2014).

****

On the Iteration Complexity of Support Recovery via Hard Thresholding Pursuit

Jie Shen, Ping Li

Recovering the support of a sparse signal from its compressed samples has been one of the most important problems in high dimensional statistics. In this paper, we present a novel analysis for the hard thresholding pursuit (HTP) algorithm, showing that it exactly recovers the support of an arbitrary s-sparse signal within O(sklogk) iterations via a properly chosen proxy function, where k is the condition number of the problem. In stark contrast to the theoretical results in the literature, the iteration complexity we obtained holds without assuming the restricted isometry property, or relaxing the sparsity, or utilizing the optimality of the underlying signal. We further extend our result to a more challenging scenario, where the subproblem involved in HTP cannot be solved exactly. We prove that even in this setting, support recovery is possible and the computational complexity of HTP is established. Numerical study substantiates our theoretical results.

****

GSOS: Gauss-Seidel Operator Splitting Algorithm for Multi-Term Nonsmooth Convex Composite Optimization

Li Shen, Wei Liu, Ganzhao Yuan, Shiqian Ma

In this paper, we propose a fast Gauss-Seidel Operator Splitting (GSOS) algorithm for addressing multi-term nonsmooth convex composite optimization, which has wide applications in machine learning, signal processing and statistics. The proposed GSOS algorithm inherits the advantage of the Gauss-Seidel technique to accelerate the optimization procedure, and leverages the operator splitting technique to reduce the computational complexity. In addition, we develop a new technique to establish the global convergence of the GSOS algorithm. To be specific, we first reformulate the iterations of GSOS as a two-step iterations algorithm by employing the tool of operator optimization theory. Subsequently, we establish the convergence of GSOS based on the two-step iterations algorithm reformulation. At last, we apply the proposed GSOS algorithm to solve overlapping group Lasso and graph-guided fused Lasso problems. Numerical experiments show that our proposed GSOS algorithm is superior to the state-of-the-art algorithms in terms of both efficiency and effectiveness.

****

World of Bits: An Open-Domain Platform for Web-Based Agents

Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, Percy Liang

While simulated game environments have greatly accelerated research in reinforcement learning, existing environments lack the open-domain realism of tasks in computer vision or natural language processing, which operate on artifacts created by humans in natural, organic settings. To foster reinforcement learning research in such settings, we introduce the World of Bits (WoB), a platform in which agents complete tasks on the Internet by performing low-level keyboard and mouse actions. The two main challenges are: (i) to curate a large, diverse set of interesting web-based tasks, and (ii) to ensure that these tasks have a well-defined reward structure and are reproducible despite the transience of the web. To do this, we develop a methodology in which crowdworkers create tasks defined by natural language questions and provide demonstrations of how to answer the question on real websites using keyboard and mouse; HTTP traffic is cached to create a reproducible offline approximation of the web site. Finally, we show that agents trained via behavioral cloning and reinforcement learning can successfully complete a range of our web-based tasks.

****

Learning Important Features Through Propagating Activation Differences
Avanti Shrikumar, Peyton Greenside, Anshul Kundaje

The purported “black box” nature of neural networks is a barrier to adoption in applications where interpretability is essential. Here we present DeepLIFT (Deep Learning Important FeaTures), a method for decomposing the output prediction of a neural network on a specific input by backpropagating the contributions of all neurons in the network to every feature of the input. DeepLIFT compares the activation of each neuron to its `reference activation’ and assigns contribution scores according to the difference. By optionally giving separate consideration to positive and negative contributions, DeepLIFT can also reveal dependencies which are missed by other approaches. Scores can be computed efficiently in a single backward pass. We apply DeepLIFT to models trained on MNIST and simulated genomic data, and show significant advantages over gradient-based methods. Video tutorial: http://goo.gl/qKb7pL code: http://goo.gl/RM8jvH

****

Optimal Densification for Fast and Accurate Minwise Hashing
Anshumali Shrivastava

Minwise hashing is a fundamental and one of the most successful hashing algorithm in the literature. Recent advances based on the idea of densification (Shrivastava \& Li, 2014) have shown that it is possible to compute k minwise hashes, of a vector with d nonzeros, in mere $(d + k)$ computations, a significant improvement over the classical $O(dk)$. These advances have led to an algorithmic improvement in the query complexity of traditional indexing algorithms based on minwise hashing. Unfortunately, the variance of the current densification techniques is unnecessarily high, which leads to significantly poor accuracy compared to vanilla minwise hashing, especially when the data is sparse. In this paper, we provide a novel densification scheme which relies on carefully tailored 2-universal hashes. We show that the proposed scheme is variance-optimal, and without losing the runtime efficiency, it is significantly more accurate than existing densification techniques. As a result, we obtain a significantly efficient hashing scheme which has the same variance and collision probability as minwise hashing. Experimental evaluations on real sparse and high-dimensional datasets validate our claims. We believe that given the significant advantages, our method will replace minwise hashing implementations in practice.

****

Bottleneck Conditional Density Estimation
Rui Shu, Hung H. Bui, Mohammad Ghavamzadeh

We introduce a new framework for training deep generative models for high-dimensional conditional density estimation. The Bottleneck Conditional Density Estimator (BCDE) is a variant of the conditional variational autoencoder (CVAE) that employs layer(s) of stochastic variables as the bottleneck between the input x and target y, where both are high-dimensional. Crucially, we propose a new hybrid training method that blends the conditional generative model with a joint generative model. Hybrid blending is the key to effective training of the BCDE, which a

voids overfitting and provides a novel mechanism for leveraging unlabeled data. We show that our hybrid training procedure enables models to achieve competitive results in the MNIST quadrant prediction task in the fully-supervised setting, and sets new benchmarks in the semi-supervised regime for MNIST, SVHN, and CelebA.
****
Attentive Recurrent Comparators
Pranav Shyam, Shubham Gupta, Ambedkar Dukkipati
Rapid learning requires flexible representations to quickly adopt to new evidence. We develop a novel class of models called Attentive Recurrent Comparators (ARCs) that form representations of objects by cycling through them and making observations. Using the representations extracted by ARCs, we develop a way of approximating a dynamic representation space and use it for one-shot learning. In the task of one-shot classification on the Omniglot dataset, we achieve the state of the art performance with an error rate of 1.5\%. This represents the first super-human result achieved for this task with a generic model that uses only pixel information.
****
Gradient Boosted Decision Trees for High Dimensional Sparse Output
Si Si, Huan Zhang, S. Sathiya Keerthi, Dhruv Mahajan, Inderjit S. Dhillon, Cho-Jui Hsieh
In this paper, we study the gradient boosted decision trees (GBDT) when the output space is high dimensional and sparse. For example, in multilabel classification, the output space is a L-dimensional 0/1 vector, where L is number of labels that can grow to millions and beyond in many modern applications. We show that vanilla GBDT can easily run out of memory or encounter near-forever running time in this regime, and propose a new GBDT variant, GBDT-SPARSE, to resolve this problem by employing L_0 regularization. We then discuss in detail how to utilize this sparsity to conduct GBDT training, including splitting the nodes, computing the sparse residual, and predicting in sublinear time. Finally, we apply our algorithm to extreme multilabel classification problems, and show that the proposed GBDT-SPARSE achieves an order of magnitude improvements in model size and prediction time over existing methods, while yielding similar performance.
****
The Predictron: End-To-End Learning and Planning
David Silver, Hado Hasselt, Matteo Hessel, Tom Schaul, Arthur Guez, Tim Harley, Gabriel Dulac-Arnold, David Reichert, Neil Rabinowitz, Andre Barreto, Thomas Degris
One of the key challenges of artificial intelligence is to learn models that are effective in the context of planning. In this document we introduce the predictron architecture. The predictron consists of a fully abstract model, represented by a Markov reward process, that can be rolled forward multiple “imagined” planning steps. Each forward pass of the predictron accumulates internal rewards and values over multiple planning depths. The predictron is trained end-to-end so as to make these accumulated values accurately approximate the true value function. We applied the predictron to procedurally generated random mazes and a simulator for the game of pool. The predictron yielded significantly more accurate predictions than conventional deep neural network architectures.
****
Fractional Langevin Monte Carlo: Exploring Levy Driven Stochastic Differential Equations for Markov Chain Monte Carlo
Umut ■im■ekli
Along with the recent advances in scalable Markov Chain Monte Carlo methods, sampling techniques that are based on Langevin diffusions have started receiving increasing attention. These so called Langevin Monte Carlo (LMC) methods are based on diffusions driven by a Brownian motion, which gives rise to Gaussian proposal distributions in the resulting algorithms. Even though these approaches have proven successful in many applications, their performance can be limited by the light-tailed nature of the Gaussian proposals. In this study, we extend classical LMC and develop a novel Fractional LMC (FLMC) framework that is based on a fami

ly of heavy-tailed distributions, called alpha-stable Levy distributions. As opposed to classical approaches, the proposed approach can possess large jumps while targeting the correct distribution, which would be beneficial for efficient exploration of the state space. We develop novel computational methods that can scale up to large-scale problems and we provide formal convergence analysis of the proposed scheme. Our experiments support our theory: FLMC can provide superior performance in multi-modal settings, improved convergence rates, and robustness to algorithm parameters.

****

Nonparanormal Information Estimation

Shashank Singh, Barnabás Póczos

We study the problem of using i.i.d. samples from an unknown multivariate probability distribution p to estimate the mutual information of p. This problem has recently received attention in two settings: (1) where p is assumed to be Gaussian and (2) where p is assumed only to lie in a large nonparametric smoothness class. Estimators proposed for the Gaussian case converge in high dimensions when the Gaussian assumption holds, but are brittle, failing dramatically when p is not Gaussian, while estimators proposed for the nonparametric case fail to converge with realistic sample sizes except in very low dimension. Hence, there is a lack of robust mutual information estimators for many realistic data. To address this, we propose estimators for mutual information when p is assumed to be a nonparanormal (or Gaussian copula) model, a semiparametric compromise between Gaussian and nonparametric extremes. Using theoretical bounds and experiments, we show these estimators strike a practical balance between robustness and scalability.

****

High-Dimensional Structured Quantile Regression

Vidyashankar Sivakumar, Arindam Banerjee

Quantile regression aims at modeling the conditional median and quantiles of a response variable given certain predictor variables. In this work we consider the problem of linear quantile regression in high dimensions where the number of predictor variables is much higher than the number of samples available for parameter estimation. We assume the true parameter to have some structure characterized as having a small value according to some atomic norm R(.) and consider the norm regularized quantile regression estimator. We characterize the sample complexity for consistent recovery and give non-asymptotic bounds on the estimation error. While this problem has been previously considered, our analysis reveals geometric and statistical characteristics of the problem not available in prior literature. We perform experiments on synthetic data which support the theoretical results.

****

Robust Budget Allocation via Continuous Submodular Functions

Matthew Staib, Stefanie Jegelka

The optimal allocation of resources for maximizing influence, spread of information or coverage, has gained attention in the past years, in particular in machine learning and data mining. But in applications, the parameters of the problem are rarely known exactly, and using wrong parameters can lead to undesirable outcomes. We hence revisit a continuous version of the Budget Allocation or Bipartite Influence Maximization problem introduced by Alon et al. (2012) from a robust optimization perspective, where an adversary may choose the least favorable parameters within a confidence set. The resulting problem is a nonconvex-concave saddle point problem (or game). We show that this nonconvex problem can be solved exactly by leveraging connections to continuous submodular functions, and by solving a constrained submodular minimization problem. Although constrained submodular minimization is hard in general, here, we establish conditions under which such a problem can be solved to arbitrary precision ϵ.

****

Probabilistic Submodular Maximization in Sub-Linear Time

Serban Stan, Morteza Zadimoghaddam, Andreas Krause, Amin Karbasi

In this paper, we consider optimizing submodular functions that are drawn from some unknown distribution. This setting arises, e.g., in recommender systems, whe re the utility of a subset of items may depend on a user-specific submodular utility function. In modern applications, the ground set of items is often so large that even the widely used (lazy) greedy algorithm is not efficient enough. As a remedy, we introduce the problem of sublinear time probabilistic submodular maximization: Given training examples of functions (e.g., via user feature vectors), we seek to reduce the ground set so that optimizing new functions drawn from the same distribution will provide almost as much value when restricted to the reduced ground set as when using the full set. We cast this problem as a two-stage submodular maximization and develop a novel efficient algorithm for this problem which offers $1/2(1 - 1/e^2)$ approximation ratio for general monotone submodular functions and general matroid constraints. We demonstrate the effectiveness of our approach on several real-world applications where running the maximization problem on the reduced ground set leads to two orders of magnitude speed-up while incurring almost no loss.

****

Approximate Steepest Coordinate Descent

Sebastian U. Stich, Anant Raj, Martin Jaggi

We propose a new selection rule for the coordinate selection in coordinate descent methods for huge-scale optimization. The efficiency of this novel scheme is provably better than the efficiency of uniformly random selection, and can reach the efficiency of steepest coordinate descent (SCD), enabling an acceleration of a factor of up to n, the number of coordinates. In many practical applications, our scheme can be implemented at no extra cost and computational efficiency very close to the faster uniform selection. Numerical experiments with Lasso and Ridge regression show promising improvements, in line with our theoretical guarantees.

****

Ordinal Graphical Models: A Tale of Two Approaches

Arun Sai Suggala, Eunho Yang, Pradeep Ravikumar

Undirected graphical models or Markov random fields (MRFs) are widely used for modeling multivariate probability distributions. Much of the work on MRFs has focused on continuous variables, and nominal variables (that is, unordered categorical variables). However, data from many real world applications involve ordered categorical variables also known as ordinal variables, e.g., movie ratings on Netflix which can be ordered from 1 to 5 stars. With respect to univariate ordinal distributions, as we detail in the paper, there are two main categories of distributions; while there have been efforts to extend these to multivariate ordinal distributions, the resulting distributions are typically very complex, with either a large number of parameters, or with non-convex likelihoods. While there have been some work on tractable approximations, these do not come with strong statistical guarantees, and moreover are relatively computationally expensive. In this paper, we theoretically investigate two classes of graphical models for ordinal data, corresponding to the two main categories of univariate ordinal distributions. In contrast to previous work, our theoretical developments allow us to provide correspondingly two classes of estimators that are not only computationally efficient but also have strong statistical guarantees.

****

Tensor Balancing on Statistical Manifold

Mahito Sugiyama, Hiroyuki Nakahara, Koji Tsuda

We solve tensor balancing, rescaling an Nth order nonnegative tensor by multiplying N tensors of order N - 1 so that every fiber sums to one. This generalizes a fundamental process of matrix balancing used to compare matrices in a wide range of applications from biology to economics. We present an efficient balancing algorithm with quadratic convergence using Newton’s method and show in numerical experiments that the proposed algorithm is several orders of magnitude faster than existing ones. To theoretically prove the correctness of the algorithm, we model tensors as probability distributions in a statistical manifold and realize tensor balancing as projection onto a submanifold. The key to our algorithm is that the gradient of the manifold, used as a Jacobian matrix in Newton’s method, can be analytically obtained using the Möbius inversion formula, the essential of combinatorial mathematics. Our model is not limited to tensor balancing, but has a wide applicability as it includes various statistical and machine learning models such as weighted DAGs and Boltzmann machines.

****

Safety-Aware Algorithms for Adversarial Contextual Bandit

Wen Sun, Debadeepta Dey, Ashish Kapoor

In this work we study the safe sequential decision making problem under the setting of adversarial contextual bandits with sequential risk constraints. At each round, nature prepares a context, a cost for each arm, and additionally a risk for each arm. The learner leverages the context to pull an arm and receives the corresponding cost and risk associated with the pulled arm. In addition to minimizing the cumulative cost, for safety purposes, the learner needs to make safe decisions such that the average of the cumulative risk from all pulled arms should not be larger than a pre-defined threshold. To address this problem, we first study online convex programming in the full information setting where in each round the learner receives an adversarial convex loss and a convex constraint. We develop a meta algorithm leveraging online mirror descent for the full information setting and then extend it to contextual bandit with sequential risk constraints setting using expert advice. Our algorithms can achieve near-optimal regret in terms of minimizing the total cost, while successfully maintaining a sub-linear growth of accumulative risk constraint violation. We support our theoretical results by demonstrating our algorithm on a simple simulated robotics reactive control task.

****

Relative Fisher Information and Natural Gradient for Learning Large Modular Models

Ke Sun, Frank Nielsen

Fisher information and natural gradient provided deep insights and powerful tools to artificial neural networks. However related analysis becomes more and more difficult as the learner’s structure turns large and complex. This paper makes a preliminary step towards a new direction. We extract a local component from a large neural system, and define its relative Fisher information metric that describes accurately this small component, and is invariant to the other parts of the system. This concept is important because the geometry structure is much simplified and it can be easily applied to guide the learning of neural networks. We provide an analysis on a list of commonly used components, and demonstrate how to use this concept to further improve optimization.

****

meProp: Sparsified Back Propagation for Accelerated Deep Learning with Reduced Overfitting

Xu Sun, Xuancheng Ren, Shuming Ma, Houfeng Wang

We propose a simple yet effective technique for neural network learning. The forward propagation is computed as usual. In back propagation, only a small subset of the full gradient is computed to update the model parameters. The gradient vectors are sparsified in such a way that only the top-k elements (in terms of magnitude) are kept. As a result, only k rows or columns (depending on the layout) of the weight matrix are modified, leading to a linear reduction (k divided by the vector dimension) in the computational cost. Surprisingly, experimental results demonstrate that we can update only 1–4\% of the weights at each back propagation pass. This does not result in a larger number of training iterations. More interestingly, the accuracy of the resulting models is actually improved rather than degraded, and a detailed analysis is given.

****

Deeply AggreVaTeD: Differentiable Imitation Learning for Sequential Prediction

Wen Sun, Arun Venkatraman, Geoffrey J. Gordon, Byron Boots, J. Andrew Bagnell

Recently, researchers have demonstrated state-of-the-art performance on sequential prediction problems using deep neural networks and Reinforcement Learning (RL). For some of these problems, oracles that can demonstrate good performance may be available during training, but are not used by plain RL methods. To take advantage of this extra information, we propose AggreVaTeD, an extension of the Imi tation Learning (IL) approach of Ross \& Bagnell (2014). AggreVaTeD allows us to use expressive differentiable policy representations such as deep networks, while leveraging training-time oracles to achieve faster and more accurate solutions with less training data. Specifically, we present two gradient procedures that can learn neural network policies for several problems, including a sequential prediction task and several high-dimensional robotics control problems. We also provide a comprehensive theoretical study of IL that demonstrates that we can expect up to exponentially-lower sample complexity for learning with AggreVaTeD than with plain RL algorithms. Our results and theory indicate that IL (and AggreVaTeD in particular) can be a more effective strategy for sequential prediction than plain RL.

****

Axiomatic Attribution for Deep Networks

Mukund Sundararajan, Ankur Taly, Qiqi Yan

We study the problem of attributing the prediction of a deep network to its input features, a problem previously studied by several other works. We identify two fundamental axioms—Sensitivity and Implementation Invariance that attribution methods ought to satisfy. We show that they are not satisfied by most known attribution methods, which we consider to be a fundamental weakness of those methods. We use the axioms to guide the design of a new attribution method called Integrated Gradients. Our method requires no modification to the original network and is extremely simple to implement; it just needs a few calls to the standard gradient operator. We apply this method to a couple of image models, a couple of text models and a chemistry model, demonstrating its ability to debug networks, to extract rules from a network, and to enable users to engage with models better.

****

Distributed Mean Estimation with Limited Communication

Ananda Theertha Suresh, Felix X. Yu, Sanjiv Kumar, H. Brendan McMahan

Motivated by the need for distributed learning and optimization algorithms with low communication cost, we study communication efficient algorithms for distributed mean estimation. Unlike previous works, we make no probabilistic assumptions on the data. We first show that for d dimensional data with n clients, a naive stochastic rounding approach yields a mean squared error (MSE) of $\Theta(d/n)$ and uses a constant number of bits per dimension per client. We then extend this naive algorithm in two ways: we show that applying a structured random rotation before quantization reduces the error to $\mathcal{O}((\log d)/n)$ and a better coding strategy further reduces the error to $\mathcal{O}(1/n)$. We also show that the latter coding strategy is optimal up to a constant in the minimax sense i.e., it achieves the best MSE for a given communication cost. We finally demonstrate the practicality of our algorithms by applying them to distributed Lloyd’s algorithm for k-means and power iteration for PCA.

****

Selective Inference for Sparse High-Order Interaction Models

Shinya Suzumura, Kazuya Nakagawa, Yuta Umezu, Koji Tsuda, Ichiro Takeuchi

Finding statistically significant high-order interactions in predictive modeling is important but challenging task because the possible number of high-order interactions is extremely large (e.g., $> 10^{17}$). In this paper we study feature selection and statistical inference for sparse high-order interaction models. Our main contribution is to extend recently developed selective inference framework for linear models to high-order interaction models by developing a novel algorithm for efficiently characterizing the selection event for the selective inference of high-order interactions. We demonstrate the effectiveness of the proposed algorithm by applying it to an HIV drug response prediction problem.

****

Coherent Probabilistic Forecasts for Hierarchical Time Series

Souhaib Ben Taieb, James W. Taylor, Rob J. Hyndman

Many applications require forecasts for a hierarchy comprising a set of time series along with aggregates of subsets of these series. Hierarchical forecasting require not only good prediction accuracy at each level of the hierarchy, but also the coherency between different levels — the property that forecasts add up ap propriately across the hierarchy. A fundamental limitation of prior research is the focus on forecasting the mean of each time series. We consider the situation where probabilistic forecasts are needed for each series in the hierarchy, and propose an algorithm to compute predictive distributions rather than mean forecasts only. Our algorithm has the advantage of synthesizing information from different levels in the hierarchy through a sparse forecast combination and a probabilistic hierarchical aggregation. We evaluate the accuracy of our forecasting algorithm on both simulated data and large-scale electricity smart meter data. The results show consistent performance gains compared to state-of-the art methods.

****

Partitioned Tensor Factorizations for Learning Mixed Membership Models

Zilong Tan, Sayan Mukherjee

We present an efficient algorithm for learning mixed membership models when the number of variables p is much larger than the number of hidden components k. This algorithm reduces the computational complexity of state-of-the-art tensor methods, which require decomposing an $O(p^3)$ tensor, to factorizing $O(p/k)$ sub-tensors each of size $O(k^3)$. In addition, we address the issue of negative entries in the empirical method of moments based estimators. We provide sufficient conditions under which our approach has provable guarantees. Our approach obtains competitive empirical results on both simulated and real data.

****

Gradient Coding: Avoiding Stragglers in Distributed Learning

Rashish Tandon, Qi Lei, Alexandros G. Dimakis, Nikos Karampatziakis

We propose a novel coding theoretic framework for mitigating stragglers in distributed learning. We show how carefully replicating data blocks and coding across gradients can provide tolerance to failures and stragglers for synchronous Gradient Descent. We implement our schemes in python (using MPI) to run on Amazon EC2, and show how we compare against baseline approaches in running time and generalization error.

****

Gradient Projection Iterative Sketch for Large-Scale Constrained Least-Squares

Junqi Tang, Mohammad Golbabaee, Mike E. Davies

We propose a randomized first order optimization algorithm Gradient Projection Iterative Sketch (GPIS) and an accelerated variant for efficiently solving large scale constrained Least Squares (LS). We provide the first theoretical convergence analysis for both algorithms. An efficient implementation using a tailored line-search scheme is also proposed. We demonstrate our methods’ computational efficiency compared to the classical accelerated gradient method, and the variance-reduced stochastic gradient methods through numerical experiments in various large synthetic/real data sets.

****

Neural Networks and Rational Functions

Matus Telgarsky

Neural networks and rational functions efficiently approximate each other. In more detail, it is shown here that for any ReLU network, there exists a rational function of degree $O(polylog(1/\epsilon))$ which is ϵ-close, and similarly for any rational function there exists a ReLU network of size $O(polylog(1/\epsilon))$ which is ϵ-close. By contrast, polynomials need degree $\Omega(poly(1/\epsilon))$ to approximate even a single ReLU. When converting a ReLU network to a rational function as above, the hidden constants depend exponentially on the number of layers, which is shown to be tight; in other words, a compositional representation can be beneficial even for rational functions.

****

Stochastic DCA for the Large-sum of Non-convex Functions Problem and its Application to Group Variable Selection in Classification

Hoai An Le Thi, Hoai Minh Le, Duy Nhat Phan, Bach Tran

In this paper, we present a stochastic version of DCA (Difference of Convex functions Algorithm) to solve a class of optimization problems whose objective function is a large sum of non-convex functions and a regularization term. We consider the $\ell_{2,0}$ regularization to deal with the group variables selection. By exploiting the special structure of the problem, we propose an efficient DC decomposition for which the corresponding stochastic DCA scheme is very inexpensive: it only requires the projection of points onto balls that is explicitly computed. As an application, we applied our algorithm for the group variables selection in multiclass logistic regression. Numerical experiments on several benchmark datasets and synthetic datasets illustrate the efficiency of our algorithm and its superiority over well-known methods, with respect to classification accuracy, sparsity of solution as well as running time.

****

An Analytical Formula of Population Gradient for two-layered ReLU network and its Applications in Convergence and Critical Point Analysis

Yuandong Tian

In this paper, we explore theoretical properties of training a two-layered ReLU network $g(\mathbf{x}; \mathbf{w}) = \sum_{j=1}^K \sigma(\mathbf{w}_j^\top\mathbf{x})$ with centered d-dimensional spherical Gaussian input $\mathbf{x}$ (σ=ReLU). We train our network with gradient descent on $\mathbf{w}$ to mimic the output of a teacher network with the same architecture and fixed parameters $\mathbf{w}^*$. We show that its population gradient has an analytical formula, leading to interesting theoretical analysis of critical points and convergence behaviors. First, we prove that critical points outside the hyperplane spanned by the teacher parameters (“out-of-plane“) are not isolated and form manifolds, and characterize in-plane critical-point-free regions for two-ReLU case. On the other hand, convergence to $\mathbf{w}^*$ for one ReLU node is guaranteed with at least $(1-\epsilon)/2$ probability, if weights are initialized randomly with standard deviation upper-bounded by $O(\epsilon/\sqrt{d})$, in accordance with empirical practice. For network with many ReLU nodes, we prove that an infinitesimal perturbation of weight initialization results in convergence towards $\mathbf{w}^*$ (or its permutation), a phenomenon known as spontaneous symmetric-breaking (SSB) in physics. We assume no independence of ReLU activations. Simulation verifies our findings.

****

Evaluating the Variance of Likelihood-Ratio Gradient Estimators

Seiya Tokui, Issei Sato

The likelihood-ratio method is often used to estimate gradients of stochastic computations, for which baselines are required to reduce the estimation variance. Many types of baselines have been proposed, although their degree of optimality is not well understood. In this study, we establish a novel framework of gradient estimation that includes most of the common gradient estimators as special cases. The framework gives a natural derivation of the optimal estimator that can be interpreted as a special case of the likelihood-ratio method so that we can evaluate the optimal degree of practical techniques with it. It bridges the likelihood-ratio method and the reparameterization trick while still supporting discrete variables. It is derived from the exchange property of the differentiation and integration. To be more specific, it is derived by the reparameterization trick and local marginalization analogous to the local expectation gradient. We evaluate various baselines and the optimal estimator for variational learning and show that the performance of the modern estimators is close to the optimal estimator.

****

Accelerating Eulerian Fluid Simulation With Convolutional Networks

Jonathan Tompson, Kristofer Schlachter, Pablo Sprechmann, Ken Perlin

Efficient simulation of the Navier-Stokes equations for fluid flow is a long standing problem in applied mathematics, for which state-of-the-art methods require large compute resources. In this work, we propose a data-driven approach that leverages the approximation power of deep-learning with the precision of standard solvers to obtain fast and highly realistic simulations. Our method solves the incompressible Euler equations using the standard operator splitting method, in which a large sparse linear system with many free parameters must be solved. We use a Convolutional Network with a highly tailored architecture, trained using a novel unsupervised learning framework to solve the linear system. We present re al-time 2D and 3D simulations that outperform recently proposed data-driven methods; the obtained results are realistic and show good generalization properties.

****

Boosted Fitted Q-Iteration

Samuele Tosatto, Matteo Pirotta, Carlo D’Eramo, Marcello Restelli

This paper is about the study of B-FQI, an Approximated Value Iteration (AVI) algorithm that exploits a boosting procedure to estimate the action-value function in reinforcement learning problems. B-FQI is an iterative off-line algorithm that, given a dataset of transitions, builds an approximation of the optimal action-value function by summing the approximations of the Bellman residuals across all iterations. The advantage of such approach w.r.t. to other AVI methods is twofold: (1) while keeping the same function space at each iteration, B-FQI can represent more complex functions by considering an additive model; (2) since the Bellman residual decreases as the optimal value function is approached, regression problems become easier as iterations proceed. We study B-FQI both theoretically, providing also a finite-sample error upper bound for it, and empirically, by comparing its performance to the one of FQI in different domains and using different regression techniques.

****

Diameter-Based Active Learning

Christopher Tosh, Sanjoy Dasgupta

To date, the tightest upper and lower-bounds for the active learning of general concept classes have been in terms of a parameter of the learning problem called the splitting index. We provide, for the first time, an efficient algorithm that is able to realize this upper bound, and we empirically demonstrate its good performance.

****

Magnetic Hamiltonian Monte Carlo

Nilesh Tripuraneni, Mark Rowland, Zoubin Ghahramani, Richard Turner

Hamiltonian Monte Carlo (HMC) exploits Hamiltonian dynamics to construct efficient proposals for Markov chain Monte Carlo (MCMC). In this paper, we present a generalization of HMC which exploits non-canonical Hamiltonian dynamics. We refer to this algorithm as magnetic HMC, since in 3 dimensions a subset of the dynamics map onto the mechanics of a charged particle coupled to a magnetic field. We establish a theoretical basis for the use of non-canonical Hamiltonian dynamics in MCMC, and construct a symplectic, leapfrog-like integrator allowing for the implementation of magnetic HMC. Finally, we exhibit several examples where these non-canonical dynamics can lead to improved mixing of magnetic HMC relative to ordinary HMC.

****

Know-Evolve: Deep Temporal Reasoning for Dynamic Knowledge Graphs

Rakshit Trivedi, Hanjun Dai, Yichen Wang, Le Song

The availability of large scale event data with time stamps has given rise to dynamically evolving knowledge graphs that contain temporal information for each edge. Reasoning over time in such dynamic knowledge graphs is not yet well understood. To this end, we present Know-Evolve, a novel deep evolutionary knowledge network that learns non-linearly evolving entity representations over time. The occurrence of a fact (edge) is modeled as a multivariate point process whose intensity function is modulated by the score for that fact computed based on the learned entity embeddings. We demonstrate significantly improved performance over various relational learning approaches on two large scale real-world datasets. Further, our method effectively predicts occurrence or recurrence time of a fact which is novel compared to prior reasoning approaches in multi-relational setting.

****

Hyperplane Clustering via Dual Principal Component Pursuit

Manolis C. Tsakiris, René Vidal

State-of-the-art methods for clustering data drawn from a union of subspaces are based on sparse and low-rank representation theory and convex optimization algorithms. Existing results guaranteeing the correctness of such methods require th e dimension of the subspaces to be small relative to the dimension of the ambient space. When this assumption is violated, as is, e.g., in the case of hyperplanes, existing methods are either computationally too intensive (e.g., algebraic methods) or lack sufficient theoretical support (e.g., K-Hyperplanes or RANSAC). In this paper we provide theoretical and algorithmic contributions to the problem of clustering data from a union of hyperplanes, by extending a recent subspace learning method called Dual Principal Component Pursuit (DPCP) to the multi-hyperplane case. We give theoretical guarantees under which, the non-convex ℓ_1 problem associated with DPCP admits a unique global minimizer equal to the normal vector of the most dominant hyperplane. Inspired by this insight, we propose sequential (RANSAC-style) and iterative (K-Hyperplanes-style) hyperplane learning DPCP algorithms, which, via experiments on synthetic and real data, are shown to outperform or be competitive to the state-of-the-art.

****

Breaking Locality Accelerates Block Gauss-Seidel

Stephen Tu, Shivaram Venkataraman, Ashia C. Wilson, Alex Gittens, Michael I. Jordan, Benjamin Recht

Recent work by Nesterov and Stich (2016) showed that momentum can be used to accelerate the rate of convergence for block Gauss-Seidel in the setting where a fixed partitioning of the coordinates is chosen ahead of time. We show that this setting is too restrictive, constructing instances where breaking locality by running non-accelerated Gauss-Seidel with randomly sampled coordinates substantially outperforms accelerated Gauss-Seidel with any fixed partitioning. Motivated by this finding, we analyze the accelerated block Gauss-Seidel algorithm in the random coordinate sampling setting. Our analysis captures the benefit of acceleration with a new data-dependent parameter which is well behaved when the matrix sub-blocks are well-conditioned. Empirically, we show that accelerated Gauss-Seidel with random coordinate sampling provides speedups for large scale machine learning tasks when compared to non-accelerated Gauss-Seidel and the classical conjugate-gradient algorithm.

****

Multilabel Classification with Group Testing and Codes

Shashanka Ubaru, Arya Mazumdar

In recent years, the multiclass and mutlilabel classification problems we encounter in many applications have very large (10^3–10^6) number of classes. However, each instance belongs to only one or few classes, i.e., the label vectors are sparse. In this work, we propose a novel approach based on group testing to solve such large multilabel classification problems with sparse label vectors. We describe various group testing constructions, and advocate the use of concatenated Reed Solomon codes and unbalanced bipartite expander graphs for extreme classification problems. The proposed approach has several advantages theoretically and practically over existing popular methods. Our method operates on the binary alphabet and can utilize the well-established binary classifiers for learning. The error correction capabilities of the codes are leveraged for the first time in the learning problem to correct prediction errors. Even if a linearly growing number of classifiers mis-classify, these errors are fully corrected. We establish Hamming loss error bounds for the approach. More importantly, our method utilizes a simple prediction algorithm and does not require matrix inversion or solving optimization problems making the algorithm very inexpensive. Numerical experiments with various datasets illustrate the superior performance of our method.

****

Learning Stable Stochastic Nonlinear Dynamical Systems

Jonas Umlauft, Sandra Hirche

A data-driven identification of dynamical systems requiring only minimal prior knowledge is promising whenever no analytically derived model structure is available, e.g., from first principles in physics. However, meta-knowledge on the system’s behavior is often given and should be exploited: Stability as fundamental property is essential when the model is used for controller design or movement generation. Therefore, this paper proposes a framework for learning stable stochastic systems from data. We focus on identifying a state-dependent coefficient for m of the nonlinear stochastic model which is globally asymptotically stable according to probabilistic Lyapunov methods. We compare our approach to other state of the art methods on real-world datasets in terms of flexibility and stability.

****

Learning Determinantal Point Processes with Moments and Cycles

John Urschel, Victor-Emmanuel Brunel, Ankur Moitra, Philippe Rigollet

Determinantal Point Processes (DPPs) are a family of probabilistic models that have a repulsive behavior, and lend themselves naturally to many tasks in machine learning where returning a diverse set of objects is important. While there are fast algorithms for sampling, marginalization and conditioning, much less is known about learning the parameters of a DPP. Our contribution is twofold: (i) we establish the optimal sample complexity achievable in this problem and show that it is governed by a natural parameter, which we call the cycle sparsity; (ii) we propose a provably fast combinatorial algorithm that implements the method of moments efficiently and achieves optimal sample complexity. Finally, we give experimental results that confirm our theoretical findings.

****

Automatic Discovery of the Statistical Types of Variables in a Dataset

Isabel Valera, Zoubin Ghahramani

A common practice in statistics and machine learning is to assume that the statistical data types (e.g., ordinal, categorical or real-valued) of variables, and usually also the likelihood model, is known. However, as the availability of real-world data increases, this assumption becomes too restrictive. Data are often heterogeneous, complex, and improperly or incompletely documented. Surprisingly, despite their practical importance, there is still a lack of tools to automatically discover the statistical types of, as well as appropriate likelihood (noise) models for, the variables in a dataset. In this paper, we fill this gap by proposing a Bayesian method, which accurately discovers the statistical data types in both synthetic and real data.

****

Model-Independent Online Learning for Influence Maximization

Sharan Vaswani, Branislav Kveton, Zheng Wen, Mohammad Ghavamzadeh, Laks V. S. Lakshmanan, Mark Schmidt

We consider influence maximization (IM) in social networks, which is the problem of maximizing the number of users that become aware of a product by selecting a set of “seed” users to expose the product to. While prior work assumes a known model of information diffusion, we propose a novel parametrization that not only makes our framework agnostic to the underlying diffusion model, but also statistically efficient to learn from data. We give a corresponding monotone, submodular surrogate function, and show that it is a good approximation to the original IM objective. We also consider the case of a new marketer looking to exploit an existing social network, while simultaneously learning the factors governing information propagation. For this, we propose a pairwise-influence semi-bandit feedback model and develop a LinUCB-based bandit algorithm. Our model-independent analysis shows that our regret bound has a better (as compared to previous work) dependence on the size of the network. Experimental evaluation suggests that our framework is robust to the underlying diffusion model and can efficiently learn a near-optimal solution.

****

FeUdal Networks for Hierarchical Reinforcement Learning

Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, Koray Kavukcuoglu

We introduce FeUdal Networks (FuNs): a novel architecture for hierarchical reinforcement learning. Our approach is inspired by the feudal reinforcement learning proposal of Dayan and Hinton, and gains power and efficacy by decoupling end-to-end learning across multiple levels – allowing it to utilise different resolutions of time. Our framework employs a Manager module and a Worker module. The Manager operates at a slower time scale and sets abstract goals which are conveyed to and enacted by the Worker. The Worker generates primitive actions at every tick of the environment. The decoupled structure of FuN conveys several benefits –

in addition to facilitating very long timescale credit assignment it also encourages the emergence of sub-policies associated with different goals set by the Manager. These properties allow FuN to dramatically outperform a strong baseline agent on tasks that involve long-term credit assignment or memorisation.

****

Scalable Multi-Class Gaussian Process Classification using Expectation Propagation

Carlos Villacampa-Calvo, Daniel Hernández-Lobato

This paper describes an expectation propagation (EP) method for multi-class classification with Gaussian processes that scales well to very large datasets. In such a method the estimate of the log-marginal-likelihood involves a sum across the data instances. This enables efficient training using stochastic gradients and mini-batches. When this type of training is used, the computational cost does not depend on the number of data instances N. Furthermore, extra assumptions in the approximate inference process make the memory cost independent of N. The consequence is that the proposed EP method can be used on datasets with millions of instances. We compare empirically this method with alternative approaches that approximate the required computations using variational inference. The results show that it performs similar or even better than these techniques, which sometimes give significantly worse predictive distributions in terms of the test log-likelihood. Besides this, the training process of the proposed approach also seems to converge in a smaller number of iterations.

****

Learning to Generate Long-term Future via Hierarchical Prediction

Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, Honglak Lee

We propose a hierarchical approach for making long-term predictions of future frames. To avoid inherent compounding errors in recursive pixel-level prediction, we propose to first estimate high-level structure in the input frames, then predict how that structure evolves in the future, and finally by observing a single frame from the past and the predicted high-level structure, we construct the future frames without having to observe any of the pixel-level predictions. Long-term video prediction is difficult to perform by recurrently observing the predicted frames because the small errors in pixel space exponentially amplify as predictions are made deeper into the future. Our approach prevents pixel-level error propagation from happening by removing the need to observe the predicted frames. Our model is built with a combination of LSTM and analogy based encoder-decoder convolutional neural networks, which independently predict the video structure and generate the future frames, respectively. In experiments, our model is evaluated on the Human3.6M and Penn Action datasets on the task of long-term pixel-level video prediction of humans performing actions and demonstrate significantly better results than the state-of-the-art.

****

On orthogonality and learning recurrent networks with long term dependencies

Eugene Vorontsov, Chiheb Trabelsi, Samuel Kadoury, Chris Pal

It is well known that it is challenging to train deep neural networks and recurrent neural networks for tasks that exhibit long term dependencies. The vanishing or exploding gradient problem is a well known issue associated with these challenges. One approach to addressing vanishing and exploding gradients is to use either soft or hard constraints on weight matrices so as to encourage or enforce orthogonality. Orthogonal matrices preserve gradient norm during backpropagation and may therefore be a desirable property. This paper explores issues with optimization convergence, speed and gradient stability when encouraging or enforcing orthogonality. To perform this analysis, we propose a weight matrix factorization and parameterization strategy through which we can bound matrix norms and therein control the degree of expansivity induced during backpropagation. We find that hard constraints on orthogonality can negatively affect the speed of convergence and model performance.

****

Fast Bayesian Intensity Estimation for the Permanental Process

Christian J. Walder, Adrian N. Bishop

The Cox process is a stochastic process which generalises the Poisson process by letting the underlying intensity function itself be a stochastic process. In this paper we present a fast Bayesian inference scheme for the permanental process, a Cox process under which the square root of the intensity is a Gaussian process. In particular we exploit connections with reproducing kernel Hilbert spaces, to derive efficient approximate Bayesian inference algorithms based on the Laplace approximation to the predictive distribution and marginal likelihood. We obtain a simple algorithm which we apply to toy and real-world problems, obtaining orders of magnitude speed improvements over previous work.

****

Optimal and Adaptive Off-policy Evaluation in Contextual Bandits

Yu-Xiang Wang, Alekh Agarwal, Miroslav Dud■■k

We study the off-policy evaluation problem—estimating the value of a target policy using data collected by another policy—under the contextual bandit model. We consider the general (agnostic) setting without access to a consistent model of rewards and establish a minimax lower bound on the mean squared error (MSE). The bound is matched up to constants by the inverse propensity scoring (IPS) and doubly robust (DR) estimators. This highlights the difficulty of the agnostic contextual setting, in contrast with multi-armed bandits and contextual bandits with access to a consistent reward model, where IPS is suboptimal. We then propose the SWITCH estimator, which can use an existing reward model (not necessarily consistent) to achieve a better bias-variance tradeoff than IPS and DR. We prove an upper bound on its MSE and demonstrate its benefits empirically on a diverse collection of datasets, often outperforming prior work by orders of magnitude.

****

Capacity Releasing Diffusion for Speed and Locality

Di Wang, Kimon Fountoulakis, Monika Henzinger, Michael W. Mahoney, Satish Rao

Diffusions and related random walk procedures are of central importance in many areas of machine learning, data analysis, and applied mathematics. Because they spread mass agnostically at each step in an iterative manner, they can sometimes spread mass “too aggressively,” thereby failing to find the “right” clusters. We introduce a novel Capacity Releasing Diffusion (CRD) Process, which is both faster and stays more local than the classical spectral diffusion process. As an application, we use our CRD Process to develop an improved local algorithm for graph clustering. Our local graph clustering method can find local clusters in a model of clustering where one begins the CRD Process in a cluster whose vertices are connected better internally than externally by an $O(\log^2 n)$ factor, where n is the number of nodes in the cluster. Thus, our CRD Process is the first local graph clustering algorithm that is not subject to the well-known quadratic Cheeger barrier. Our result requires a certain smoothness condition, which we expect to be an artifact of our analysis. Our empirical evaluation demonstrates improved results, in particular for realistic social graphs where there are moderately good—but not very good—clusters.

****

Sketched Ridge Regression: Optimization Perspective, Statistical Perspective, and Model Averaging

Shusen Wang, Alex Gittens, Michael W. Mahoney

We address the statistical and optimization impacts of using classical sketch versus Hessian sketch to solve approximately the Matrix Ridge Regression (MRR) problem. Prior research has considered the effects of classical sketch on least squares regression (LSR), a strictly simpler problem. We establish that classical sketch has a similar effect upon the optimization properties of MRR as it does on those of LSR—namely, it recovers nearly optimal solutions. In contrast, Hessian sketch does not have this guarantee; instead, the approximation error is governed by a subtle interplay between the “mass” in the responses and the optimal objective value. For both types of approximations, the regularization in the sketched MRR problem gives it significantly different statistical properties from the sketched LSR problem. In particular, there is a bias-variance trade-off in sketched MRR that is not present in sketched LSR. We provide upper and lower bounds on the biases and variances of sketched MRR; these establish that the variance is significantly increased when classical sketches are used, while the bias is significantly increased when using Hessian sketches. Empirically, sketched MRR solutions can have risks that are higher by an order-of-magnitude than those of the optimal MRR solutions. We establish theoretically and empirically that model averaging greatly decreases this gap. Thus, in the distributed setting, sketching combined with model averaging is a powerful technique that quickly obtains near-optimal solutions to the MRR problem while greatly mitigating the statistical risks incurred by sketching.

****

Robust Gaussian Graphical Model Estimation with Arbitrary Corruption

Lingxiao Wang, Quanquan Gu

We study the problem of estimating the high-dimensional Gaussian graphical model where the data are arbitrarily corrupted. We propose a robust estimator for the sparse precision matrix in the high-dimensional regime. At the core of our method is a robust covariance matrix estimator, which is based on truncated inner product. We establish the statistical guarantee of our estimator on both estimation error and model selection consistency. In particular, we show that provided that the number of corrupted samples n_2 for each variable satisfies $n_2 \lesssim \sqrt{n}/\sqrt{\log d}$, where n is the sample size and d is the number of variables, the proposed robust precision matrix estimator attains the same statistical rate as the standard estimator for Gaussian graphical models. In addition, we propose a hypothesis testing procedure to assess the uncertainty of our robust estimator. We demonstrate the effectiveness of our method through extensive experiments on both synthetic data and real-world genomic data.

****

Max-value Entropy Search for Efficient Bayesian Optimization

Zi Wang, Stefanie Jegelka

Entropy Search (ES) and Predictive Entropy Search (PES) are popular and empirically successful Bayesian Optimization techniques. Both rely on a compelling information-theoretic motivation, and maximize the information gained about the $\arg\max$ of the unknown function; yet, both are plagued by the expensive computation for estimating entropies. We propose a new criterion, Max-value Entropy Search (MES), that instead uses the information about the maximum function value. We show relations of MES to other Bayesian optimization methods, and establish a regret bound. We observe that MES maintains or improves the good empirical performance of ES/PES, while tremendously lightening the computational burden. In particular, MES is much more robust to the number of samples used for computing the entropy, and hence more efficient for higher dimensional problems.

****

Efficient Distributed Learning with Sparsity

Jialei Wang, Mladen Kolar, Nathan Srebro, Tong Zhang

We propose a novel, efficient approach for distributed sparse learning with observations randomly partitioned across machines. In each round of the proposed method, worker machines compute the gradient of the loss on local data and the master machine solves a shifted ℓ_1 regularized loss minimization problem. After a number of communication rounds that scales only logarithmically with the number of machines, and independent of other parameters of the problem, the proposed approach provably matches the estimation error bound of centralized methods.

****

Robust Probabilistic Modeling with Bayesian Data Reweighting

Yixin Wang, Alp Kucukelbir, David M. Blei

Probabilistic models analyze data by relying on a set of assumptions. Data that exhibit deviations from these assumptions can undermine inference and prediction quality. Robust models offer protection against mismatch between a model’s assumptions and reality. We propose a way to systematically detect and mitigate mismatch of a large class of probabilistic models. The idea is to raise the likelihood of each observation to a weight and then to infer both the latent variables and the weights from data. Inferring the weights allows a model to identify observations that match its assumptions and down-weight others. This enables robust inference and improves predictive accuracy. We study four different forms of mism atch with reality, ranging from missing latent groups to structure misspecification. A Poisson factorization analysis of the Movielens 1M dataset shows the benefits of this approach in a practical scenario.

****

Batched High-dimensional Bayesian Optimization via Structural Kernel Learning

Zi Wang, Chengtao Li, Stefanie Jegelka, Pushmeet Kohli

Optimization of high-dimensional black-box functions is an extremely challenging problem. While Bayesian optimization has emerged as a popular approach for optimizing black-box functions, its applicability has been limited to low-dimensional problems due to its computational and statistical challenges arising from high-dimensional settings. In this paper, we propose to tackle these challenges by (1) assuming a latent additive structure in the function and inferring it properly for more efficient and effective BO, and (2) performing multiple evaluations in parallel to reduce the number of iterations required by the method. Our novel approach learns the latent structure with Gibbs sampling and constructs batched queries using determinantal point processes. Experimental validations on both synthetic and real-world functions demonstrate that the proposed method outperforms the existing state-of-the-art approaches.

****

Tensor Decomposition via Simultaneous Power Iteration

Po-An Wang, Chi-Jen Lu

Tensor decomposition is an important problem with many applications across several disciplines, and a popular approach for this problem is the tensor power method. However, previous works with theoretical guarantee based on this approach can only find the top eigenvectors one after one, unlike the case for matrices. In this paper, we show how to find the eigenvectors simultaneously with the help of a new initialization procedure. This allows us to achieve a better running time in the batch setting, as well as a lower sample complexity in the streaming setting.

****

Sequence Modeling via Segmentations

Chong Wang, Yining Wang, Po-Sen Huang, Abdelrahman Mohamed, Dengyong Zhou, Li Deng

Segmental structure is a common pattern in many types of sequences such as phrases in human languages. In this paper, we present a probabilistic model for sequences via their segmentations. The probability of a segmented sequence is calculated as the product of the probabilities of all its segments, where each segment is modeled using existing tools such as recurrent neural networks. Since the segmentation of a sequence is usually unknown in advance, we sum over all valid segmentations to obtain the final probability for the sequence. An efficient dynamic programming algorithm is developed for forward and backward computations without resorting to any approximation. We demonstrate our approach on text segmentation and speech recognition tasks. In addition to quantitative results, we also show that our approach can discover meaningful segments in their respective application contexts.

****

Variational Policy for Guiding Point Processes

Yichen Wang, Grady Williams, Evangelos Theodorou, Le Song

Temporal point processes have been widely applied to model event sequence data generated by online users. In this paper, we consider the problem of how to design the optimal control policy for point processes, such that the stochastic system driven by the point process is steered to a target state. In particular, we exploit the key insight to view the stochastic optimal control problem from the perspective of optimal measure and variational inference. We further propose a convex optimization framework and an efficient algorithm to update the policy adaptively to the current system state. Experiments on synthetic and real-world data show that our algorithm can steer the user activities much more accurately and efficiently than other stochastic control methods.

****

Exploiting Strong Convexity from Data with Primal-Dual First-Order Algorithms

Jialei Wang, Lin Xiao

We consider empirical risk minimization of linear predictors with convex loss functions. Such problems can be reformulated as convex-concave saddle point problems and solved by primal-dual first-order algorithms. However, primal-dual algorithms often require explicit strongly convex regularization in order to obtain fast linear convergence, and the required dual proximal mapping may not admit closed-form or efficient solution. In this paper, we develop both batch and randomized primal-dual algorithms that can exploit strong convexity from data adaptively and are capable of achieving linear convergence even without regularization. We also present dual-free variants of adaptive primal-dual algorithms that do not need the dual proximal mapping, which are especially suitable for logistic regression.

****

Beyond Filters: Compact Feature Map for Portable Deep Model

Yunhe Wang, Chang Xu, Chao Xu, Dacheng Tao

Convolutional neural networks (CNNs) have shown extraordinary performance in a number of applications, but they are usually of heavy design for the accuracy reason. Beyond compressing the filters in CNNs, this paper focuses on the redundancy in the feature maps derived from the large number of filters in a layer. We propose to extract intrinsic representation of the feature maps and preserve the discriminability of the features. Circulant matrix is employed to formulate the feature map transformation, which only requires O(dlog d) computation complexity to embed a d-dimensional feature map. The filter is then re-configured to establish the mapping from original input to the new compact feature map, and the resulting network can preserve intrinsic information of the original network with significantly fewer parameters, which not only decreases the online memory for launching CNN but also accelerates the computation speed. Experiments on benchmark image datasets demonstrate the superiority of the proposed algorithm over state-of-the-art methods.

****

A Unified Variance Reduction-Based Framework for Nonconvex Low-Rank Matrix Recovery

Lingxiao Wang, Xiao Zhang, Quanquan Gu

We propose a generic framework based on a new stochastic variance-reduced gradient descent algorithm for accelerating nonconvex low-rank matrix recovery. Starting from an appropriate initial estimator, our proposed algorithm performs projected gradient descent based on a novel semi-stochastic gradient specifically designed for low-rank matrix recovery. Based upon the mild restricted strong convexity and smoothness conditions, we derive a projected notion of the restricted Lipschitz continuous gradient property, and prove that our algorithm enjoys linear convergence rate to the unknown low-rank matrix with an improved computational complexity. Moreover, our algorithm can be employed to both noiseless and noisy observations, where the (near) optimal sample complexity and statistical rate can be attained respectively. We further illustrate the superiority of our generic framework through several specific examples, both theoretically and experimentally.

****

Source-Target Similarity Modelings for Multi-Source Transfer Gaussian Process Regression

Pengfei Wei, Ramon Sagarna, Yiping Ke, Yew-Soon Ong, Chi-Keong Goh

A key challenge in multi-source transfer learning is to capture the diverse inter-domain similarities. In this paper, we study different approaches based on Gaussian process models to solve the multi-source transfer regression problem. Precisely, we first investigate the feasibility and performance of a family of transfer covariance functions that represent the pairwise similarity of each source and the target domain. We theoretically show that using such a transfer covariance function for general Gaussian process modelling can only capture the same similarity coefficient for all the sources, and thus may result in unsatisfactory transfer performance. This leads us to propose TC$_{MS}$Stack, an integrated strategy incorporating the benefits of the transfer covariance function and stacking.

Extensive experiments on one synthetic and two real-world datasets, with learning settings of up to 11 sources for the latter, demonstrate the effectiveness of our proposed TC$_{MS}$Stack.
****

Latent Intention Dialogue Models
Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, Steve Young
Developing a dialogue agent that is capable of making autonomous decisions and communicating by natural language is one of the long-term goals of machine learning research. The traditional approaches either rely on hand-crafting a small state-action set for applying reinforcement learning that is not scalable or constructing deterministic models for learning dialogue sentences that fail to capture the conversational stochasticity. In this paper, however, we propose a Latent Intention Dialogue Model that employs a discrete latent variable to learn underlying dialogue intentions in the framework of neural variational inference. Additionally, in a goal-oriented dialogue scenario, the latent intentions can be interpreted as actions guiding the generation of machine responses, which can be further refined autonomously by reinforcement learning. The experiments demonstrate the effectiveness of discrete latent variable models on learning goal-oriented dialogues, and the results outperform the published benchmarks on both corpus-based evaluation and human evaluation.
****

Unifying Task Specification in Reinforcement Learning
Martha White
Reinforcement learning tasks are typically specified as Markov decision processes. This formalism has been highly successful, though specifications often couple the dynamics of the environment and the learning objective. This lack of modularity can complicate generalization of the task specification, as well as obfuscate connections between different task settings, such as episodic and continuing. In this work, we introduce the RL task formalism, that provides a unification through simple constructs including a generalization to transition-based discounting. Through a series of examples, we demonstrate the generality and utility of this formalism. Finally, we extend standard learning constructs, including Bellman operators, and extend some seminal theoretical results, including approximation errors bounds. Overall, we provide a well-understood and sound formalism on which to build theoretical results and simplify algorithm use and development.
****

Learned Optimizers that Scale and Generalize
Olga Wichrowska, Niru Maheswaranathan, Matthew W. Hoffman, Sergio Gómez Colmenarejo, Misha Denil, Nando Freitas, Jascha Sohl-Dickstein
Learning to learn has emerged as an important direction for achieving artificial intelligence. Two of the primary barriers to its adoption are an inability to scale to larger problems and a limited ability to generalize to new tasks. We introduce a learned gradient descent optimizer that generalizes well to new tasks, and which has significantly reduced memory and computation overhead. We achieve this by introducing a novel hierarchical RNN architecture, with minimal per-parameter overhead, augmented with additional architectural features that mirror the known structure of optimization tasks. We also develop a meta-training ensemble of small, diverse, optimization tasks capturing common properties of loss landscapes. The optimizer learns to outperform RMSProp/ADAM on problems in this corpus. More importantly, it performs comparably or better when applied to small convolutional neural networks, despite seeing no neural networks in its meta-training set. Finally, it generalizes to train Inception V3 and ResNet V2 architectures on the ImageNet dataset for thousands of steps, optimization problems that are of a vastly different scale than those it was trained on.
****

Exact Inference for Integer Latent-Variable Models
Kevin Winner, Debora Sujono, Dan Sheldon
Graphical models with latent count variables arise in a number of areas. However, standard inference algorithms do not apply to these models due to the infinite support of the latent variables. Winner and Sheldon (2016) recently developed a

new technique using probability generating functions (PGFs) to perform efficient, exact inference for certain Poisson latent variable models. However, the method relies on symbolic manipulation of PGFs, and it is unclear whether this can be extended to more general models. In this paper we introduce a new approach for inference with PGFs: instead of manipulating PGFs symbolically, we adapt techniques from the autodiff literature to compute the higher-order derivatives necessary for inference. This substantially generalizes the class of models for which efficient, exact inference algorithms are available. Specifically, our results apply to a class of models that includes branching processes, which are widely used in applied mathematics and population ecology, and autoregressive models for integer data. Experiments show that our techniques are more scalable than existing approximate methods and enable new applications.

****

Tensor Belief Propagation

Andrew Wrigley, Wee Sun Lee, Nan Ye

We propose a new approximate inference algorithm for graphical models, tensor belief propagation, based on approximating the messages passed in the junction tree algorithm. Our algorithm represents the potential functions of the graphical model and all messages on the junction tree compactly as mixtures of rank-1 tensors. Using this representation, we show how to perform the operations required for inference on the junction tree efficiently: marginalisation can be computed quickly due to the factored form of rank-1 tensors while multiplication can be approximated using sampling. Our analysis gives sufficient conditions for the algorithm to perform well, including for the case of high-treewidth graphs, for which exact inference is intractable. We compare our algorithm experimentally with several approximate inference algorithms and show that it performs well.

****

A Unified View of Multi-Label Performance Measures

Xi-Zhu Wu, Zhi-Hua Zhou

Multi-label classification deals with the problem where each instance is associated with multiple class labels. Because evaluation in multi-label classification is more complicated than single-label setting, a number of performance measures have been proposed. It is noticed that an algorithm usually performs differently on different measures. Therefore, it is important to understand which algorithms perform well on which measure(s) and why. In this paper, we propose a unified margin view to revisit eleven performance measures in multi-label classification. In particular, we define label-wise margin and instance-wise margin, and prove that through maximizing these margins, different corresponding performance measures are to be optimized. Based on the defined margins, a max-margin approach called LIMO is designed and empirical results validate our theoretical findings.

****

Dual Supervised Learning

Yingce Xia, Tao Qin, Wei Chen, Jiang Bian, Nenghai Yu, Tie-Yan Liu

Many supervised learning tasks are emerged in dual forms, e.g., English-to-French translation vs. French-to-English translation, speech recognition vs. text to speech, and image classification vs. image generation. Two dual tasks have intrinsic connections with each other due to the probabilistic correlation between their models. This connection is, however, not effectively utilized today, since people usually train the models of two dual tasks separately and independently. In this work, we propose training the models of two dual tasks simultaneously, and explicitly exploiting the probabilistic correlation between them to regularize the training process. For ease of reference, we call the proposed approach dual supervised learning. We demonstrate that dual supervised learning can improve the practical performances of both tasks, for various applications including machine translation, image processing, and sentiment analysis.

****

Learning Latent Space Models with Angular Constraints

Pengtao Xie, Yuntian Deng, Yi Zhou, Abhimanu Kumar, Yaoliang Yu, James Zou, Eric P. Xing

The large model capacity of latent space models (LSMs) enables them to achieve g reat performance on various applications, but meanwhile renders LSMs to be prone to overfitting. Several recent studies investigate a new type of regularization approach, which encourages components in LSMs to be diverse, for the sake of alleviating overfitting. While they have shown promising empirical effectiveness, in theory why larger “diversity” results in less overfitting is still unclear. To bridge this gap, we propose a new diversity-promoting approach that is both theoretically analyzable and empirically effective. Specifically, we use near-orthogonality to characterize “diversity” and impose angular constraints (ACs) on the components of LSMs to promote diversity. A generalization error analysis shows that larger diversity results in smaller estimation error and larger approximation error. An efficient ADMM algorithm is developed to solve the constrained LSM problems. Experiments demonstrate that ACs improve generalization performance of LSMs and outperform other diversity-promoting approaches.

****

Uncorrelation and Evenness: a New Diversity-Promoting Regularizer

Pengtao Xie, Aarti Singh, Eric P. Xing

Latent space models (LSMs) provide a principled and effective way to extract hidden patterns from observed data. To cope with two challenges in LSMs: (1) how to capture infrequent patterns when pattern frequency is imbalanced and (2) how to reduce model size without sacrificing their expressiveness, several studies have been proposed to “diversify” LSMs, which design regularizers to encourage the components therein to be “diverse”. In light of the limitations of existing approaches, we design a new diversity-promoting regularizer by considering two factors: uncorrelation and evenness, which encourage the components to be uncorrelated and to play equally important roles in modeling data. Formally, this amounts to encouraging the covariance matrix of the components to have more uniform eigenvalues. We apply the regularizer to two LSMs and develop an efficient optimization algorithm. Experiments on healthcare, image and text data demonstrate the effectiveness of the regularizer.

****

Stochastic Convex Optimization: Faster Local Growth Implies Faster Global Convergence

Yi Xu, Qihang Lin, Tianbao Yang

In this paper, a new theory is developed for first-order stochastic convex optimization, showing that the global convergence rate is sufficiently quantified by a local growth rate of the objective function in a neighborhood of the optimal solutions. In particular, if the objective function $F(\mathbf{w})$ in the ϵ-sublevel set grows as fast as $\|\mathbf{w} - \mathbf{w}_*\|_2^{1/\theta}$, where $\mathbf{w}_*$ represents the closest optimal solution to $\mathbf{w}$ and $\theta\in(0,1]$ quantifies the local growth rate, the iteration complexity of first-order stochastic optimization for achieving an ϵ-optimal solution can be $\widetilde O(1/\epsilon^{2(1-\theta)})$, which is optimal at most up to a logarithmic factor. This result is fundamentally better in contrast with the previous works that either assume a global growth condition in the entire domain or achieve a local faster convergence under the local faster growth condition. To achieve the faster global convergence, we develop two different accelerated stochastic subgradient methods by iteratively solving the original problem approximately in a local region around a historical solution with the size of the local region gradually decreasing as the solution approaches the optimal set. Besides the theoretical improvements, this work also include new contributions towards making the proposed algorithms practical: (i) we present practical variants of accelerated stochastic subgradient methods that can run without the knowledge of multiplicative growth constant and even the growth rate θ; (ii) we consider a broad family of problems in machine learning to demonstrate that the proposed algorithms enjoy faster convergence than traditional stochastic subgradient method. For example, when applied to the ℓ_1 regularized empirical polyhedral loss minimization (e.g., hinge loss, absolute loss), the proposed stochastic methods have a logarithmic iteration complexity.

****

Learning Hawkes Processes from Short Doubly-Censored Event Sequences

Hongteng Xu, Dixin Luo, Hongyuan Zha

Many real-world applications require robust algorithms to learn point process models based on a type of incomplete data — the so-called short doubly-censored (SDC) event sequences. In this paper, we study this critical problem of quantitative asynchronous event sequence analysis under the framework of Hawkes processes by leveraging the general idea of data synthesis. In particular, given SDC event sequences observed in a variety of time intervals, we propose a sampling-stitching data synthesis method — sampling predecessor and successor for each SDC event sequence from potential candidates and stitching them together to synthesize long training sequences. The rationality and the feasibility of our method are discussed in terms of arguments based on likelihood. Experiments on both synthetic and real-world data demonstrate that the proposed data synthesis method improves learning results indeed for both time-invariant and time-varying Hawkes processes.

****

Adaptive Consensus ADMM for Distributed Optimization

Zheng Xu, Gavin Taylor, Hao Li, Mário A. T. Figueiredo, Xiaoming Yuan, Tom Goldstein

The alternating direction method of multipliers (ADMM) is commonly used for distributed model fitting problems, but its performance and reliability depend strongly on user-defined penalty parameters. We study distributed ADMM methods that boost performance by using different fine-tuned algorithm parameters on each worker node. We present a O(1/k) convergence rate for adaptive ADMM methods with node-specific parameters, and propose adaptive consensus ADMM (ACADMM), which automatically tunes parameters without user oversight.

****

High-dimensional Non-Gaussian Single Index Models via Thresholded Score Function Estimation

Zhuoran Yang, Krishnakumar Balasubramanian, Han Liu

We consider estimating the parametric component of single index models in high dimensions. Compared with existing work, we do not require the covariate to be normally distributed. Utilizing Stein’s Lemma, we propose estimators based on the score function of the covariate. Moreover, to handle score function and response variables that are heavy-tailed, our estimators are constructed via carefully thresholding their empirical counterparts. Under a bounded fourth moment condition, we establish optimal statistical rates of convergence for the proposed estimators. Extensive numerical experiments are provided to back up our theory.

****

Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering

Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, Mingyi Hong

Most learning approaches treat dimensionality reduction (DR) and clustering separately (i.e., sequentially), but recent research has shown that optimizing the two tasks jointly can substantially improve the performance of both. The premise behind the latter genre is that the data samples are obtained via linear transformation of latent representations that are easy to cluster; but in practice, the transformation from the latent space to the data can be more complicated. In this work, we assume that this transformation is an unknown and possibly nonlinear function. To recover the `clustering-friendly’ latent representations and to better cluster the data, we propose a joint DR and K-means clustering approach in which DR is accomplished via learning a deep neural network (DNN). The motivation is to keep the advantages of jointly optimizing the two tasks, while exploiting the deep neural network’s ability to approximate any nonlinear function. This way, the proposed approach can work well for a broad class of generative models. Towards this end, we carefully design the DNN structure and the associated joint optimization criterion, and propose an effective and scalable algorithm to handle the formulated optimization problem. Experiments using different real datasets are employed to showcase the effectiveness of the proposed approach.

****

On The Projection Operator to A Three-view Cardinality Constrained Set

Haichuan Yang, Shupeng Gui, Chuyang Ke, Daniel Stefankovic, Ryohei Fujimaki, Ji

Liu

The cardinality constraint is an intrinsic way to restrict the solution structure in many domains, for example, sparse learning, feature selection, and compressed sensing. To solve a cardinality constrained problem, the key challenge is to solve the projection onto the cardinality constraint set, which is NP-hard in general when there exist multiple overlapped cardinality constraints. In this paper, we consider the scenario where the overlapped cardinality constraints satisfy a Three-view Cardinality Structure (TVCS), which reflects the natural restriction in many applications, such as identification of gene regulatory networks and task-worker assignment problem. We cast the projection into a linear programming, and show that for TVCS, the vertex solution of this linear programming is the solution for the original projection problem. We further prove that such solution can be found with the complexity proportional to the number of variables and constraints. We finally use synthetic experiments and two interesting applications in bioinformatics and crowdsourcing to validate the proposed TVCS model and method.

****

Improved Variational Autoencoders for Text Modeling using Dilated Convolutions

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, Taylor Berg-Kirkpatrick

Recent work on generative text modeling has found that variational autoencoders (VAE) with LSTM decoders perform worse than simpler LSTM language models (Bowman et al., 2015). This negative result is so far poorly understood, but has been attributed to the propensity of LSTM decoders to ignore conditioning information from the encoder. In this paper, we experiment with a new type of decoder for VAE: a dilated CNN. By changing the decoder’s dilation architecture, we control the size of context from previously generated words. In experiments, we find that there is a trade-off between contextual capacity of the decoder and effective use of encoding information. We show that when carefully managed, VAEs can outperform LSTM language models. We demonstrate perplexity gains on two datasets, representing the first positive language modeling result with VAE. Further, we conduct an in-depth investigation of the use of VAE (with our new decoding architecture) for semi-supervised and unsupervised labeling tasks, demonstrating gains over several strong baselines.

****

Tensor-Train Recurrent Neural Networks for Video Classification

Yinchong Yang, Denis Krompass, Volker Tresp

The Recurrent Neural Networks and their variants have shown promising performances in sequence modeling tasks such as Natural Language Processing. These models, however, turn out to be impractical and difficult to train when exposed to very high-dimensional inputs due to the large input-to-hidden weight matrix. This may have prevented RNNs’ large-scale application in tasks that involve very high input dimensions such as video modeling; current approaches reduce the input dimensions using various feature extractors. To address this challenge, we propose a new, more general and efficient approach by factorizing the input-to-hidden weight matrix using Tensor-Train decomposition which is trained simultaneously with the weights themselves. We test our model on classification tasks using multiple real-world video datasets and achieve competitive performances with state-of-the-art models, even though our model architecture is orders of magnitude less complex. We believe that the proposed approach provides a novel and fundamental building block for modeling high-dimensional sequential data with RNN architectures and opens up many possibilities to transfer the expressive and advanced architectures from other domains such as NLP to modeling high-dimensional sequential data.

****

A Richer Theory of Convex Constrained Optimization with Reduced Projections and Improved Rates

Tianbao Yang, Qihang Lin, Lijun Zhang

This paper focuses on convex constrained optimization problems, where the solution is subject to a convex inequality constraint. In particular, we aim at challenging problems for which both projection into the constrained domain and a linea r optimization under the inequality constraint are time-consuming, which render both projected gradient methods and conditional gradient methods (a.k.a. the Frank-Wolfe algorithm) expensive. In this paper, we develop projection reduced optimization algorithms for both smooth and non-smooth optimization with improved convergence rates under a certain regularity condition of the constraint function. We first present a general theory of optimization with only one projection. Its application to smooth optimization with only one projection yields $O(1/\epsilon)$ iteration complexity, which improves over the $O(1/\epsilon^2)$ iteration complexity established before for non-smooth optimization and can be further reduced under strong convexity. Then we introduce a local error bound condition and develop faster algorithms for non-strongly convex optimization at the price of a logarithmic number of projections. In particular, we achieve an iteration complexity of $\widetilde O(1/\epsilon^{2(1-\theta)})$ for non-smooth optimization and $\widetilde O(1/\epsilon^{1-\theta})$ for smooth optimization, where $\theta\in(0,1]$ appearing the local error bound condition characterizes the functional local growth rate around the optimal solutions. Novel applications in solving the constrained ℓ_1 minimization problem and a positive semi-definite constrained distance metric learning problem demonstrate that the proposed algorithms achieve significant speed-up compared with previous algorithms.

****

Sparse + Group-Sparse Dirty Models: Statistical Guarantees without Unreasonable Conditions and a Case for Non-Convexity
Eunho Yang, Aurélie C. Lozano

Imposing sparse + group-sparse superposition structures in high-dimensional parameter estimation is known to provide flexible regularization that is more realistic for many real-world problems. For example, such a superposition enables partially-shared support sets in multi-task learning, thereby striking the right balance between parameter overlap across tasks and task specificity. Existing theoretical results on estimation consistency, however, are problematic as they require too stringent an assumption: the incoherence between sparse and group-sparse superposed components. In this paper, we fill the gap between the practical success and suboptimal analysis of sparse + group-sparse models, by providing the first consistency results that do not require unrealistic assumptions. We also study non-convex counterparts of sparse + group-sparse models. Interestingly, we show that these are guaranteed to recover the true support set under much milder conditions and with smaller sample size than convex models, which might be critical in practical applications as illustrated by our experiments.

****

Scalable Bayesian Rule Lists
Hongyu Yang, Cynthia Rudin, Margo Seltzer

We present an algorithm for building probabilistic rule lists that is two orders of magnitude faster than previous work. Rule list algorithms are competitors for decision tree algorithms. They are associative classifiers, in that they are built from pre-mined association rules. They have a logical structure that is a sequence of IF-THEN rules, identical to a decision list or one-sided decision tree. Instead of using greedy splitting and pruning like decision tree algorithms, we aim to fully optimize over rule lists, striking a practical balance between accuracy, interpretability, and computational speed. The algorithm presented here uses a mixture of theoretical bounds (tight enough to have practical implications as a screening or bounding procedure), computational reuse, and highly tuned language libraries to achieve computational efficiency. Currently, for many practical problems, this method achieves better accuracy and sparsity than decision trees. In many cases, the computational time is practical and often less than that of decision trees.

****

Approximate Newton Methods and Their Local Convergence
Haishan Ye, Luo Luo, Zhihua Zhang

Many machine learning models are reformulated as optimization problems. Thus, it is important to solve a large-scale optimization problem in big data applications. Recently, subsampled Newton methods have emerged to attract much attention f or optimization due to their efficiency at each iteration, rectified a weakness in the ordinary Newton method of suffering a high cost in each iteration while commanding a high convergence rate. Other efficient stochastic second order methods are also proposed. However, the convergence properties of these methods are still not well understood. There are also several important gaps between the current convergence theory and performance in real applications. In this paper, we aim to fill these gaps. We propose a unifying framework to analyze local convergence properties of second order methods. Based on this framework, our theoretical analysis matches the performance in real applications.

****

A Simulated Annealing Based Inexact Oracle for Wasserstein Loss Minimization
Jianbo Ye, James Z. Wang, Jia Li

Learning under a Wasserstein loss, a.k.a. Wasserstein loss minimization (WLM), is an emerging research topic for gaining insights from a large set of structured objects. Despite being conceptually simple, WLM problems are computationally challenging because they involve minimizing over functions of quantities (i.e. Wasserstein distances) that themselves require numerical algorithms to compute. In this paper, we introduce a stochastic approach based on simulated annealing for solving WLMs. Particularly, we have developed a Gibbs sampler to approximate effectively and efficiently the partial gradients of a sequence of Wasserstein losses. Our new approach has the advantages of numerical stability and readiness for warm starts. These characteristics are valuable for WLM problems that often require multiple levels of iterations in which the oracle for computing the value and gradient of a loss function is embedded. We applied the method to optimal transport with Coulomb cost and the Wasserstein non-negative matrix factorization problem, and made comparisons with the existing method of entropy regularization.

****

Latent Feature Lasso
Ian En-Hsu Yen, Wei-Cheng Lee, Sung-En Chang, Arun Sai Suggala, Shou-De Lin, Pradeep Ravikumar

The latent feature model (LFM), proposed in (Griffiths \& Ghahramani, 2005), but possibly with earlier origins, is a generalization of a mixture model, where each instance is generated not from a single latent class but from a combination of latent features. Thus, each instance has an associated latent binary feature incidence vector indicating the presence or absence of a feature. Due to its combinatorial nature, inference of LFMs is considerably intractable, and accordingly, most of the attention has focused on nonparametric LFMs, with priors such as the Indian Buffet Process (IBP) on infinite binary matrices. Recent efforts to tackle this complexity either still have computational complexity that is exponential, or sample complexity that is high-order polynomial w.r.t. the number of latent features. In this paper, we address this outstanding problem of tractable estimation of LFMs via a novel atomic-norm regularization, which gives an algorithm with polynomial run-time and sample complexity without impractical assumptions on the data distribution.

****

Combined Group and Exclusive Sparsity for Deep Neural Networks
Jaehong Yoon, Sung Ju Hwang

The number of parameters in a deep neural network is usually very large, which helps with its learning capacity but also hinders its scalability and practicality due to memory/time inefficiency and overfitting. To resolve this issue, we propose a sparsity regularization method that exploits both positive and negative correlations among the features to enforce the network to be sparse, and at the same time remove any redundancies among the features to fully utilize the capacity of the network. Specifically, we propose to use an exclusive sparsity regularization based on (1,2)-norm, which promotes competition for features between different weights, thus enforcing them to fit to disjoint sets of features. We further combine the exclusive sparsity with the group sparsity based on (2,1)-norm, to promote both sharing and competition for features in training of a deep neural network. We validate our method on multiple public datasets, and the results show that our method can obtain more compact and efficient networks while also imp roving the performance over the base networks with full weights, as opposed to existing sparsity regularizations that often obtain efficiency at the expense of prediction accuracy.

****

Latent LSTM Allocation: Joint Clustering and Non-Linear Dynamic Modeling of Sequence Data

Manzil Zaheer, Amr Ahmed, Alexander J. Smola

Recurrent neural networks, such as long-short term memory (LSTM) networks, are powerful tools for modeling sequential data like user browsing history (Tan et al., 2016; Korpusik et al., 2016) or natural language text (Mikolov et al., 2010). However, to generalize across different user types, LSTMs require a large number of parameters, notwithstanding the simplicity of the underlying dynamics, rendering it uninterpretable, which is highly undesirable in user modeling. The increase in complexity and parameters arises due to a large action space in which many of the actions have similar intent or topic. In this paper, we introduce Latent LSTM Allocation (LLA) for user modeling combining hierarchical Bayesian models with LSTMs. In LLA, each user is modeled as a sequence of actions, and the model jointly groups actions into topics and learns the temporal dynamics over the topic sequence, instead of action space directly. This leads to a model that is highly interpretable, concise, and can capture intricate dynamics. We present an efficient Stochastic EM inference algorithm for our model that scales to millions of users/documents. Our experimental evaluations show that the proposed model compares favorably with several state-of-the-art baselines.

****

Canopy Fast Sampling with Cover Trees

Manzil Zaheer, Satwik Kottur, Amr Ahmed, José Moura, Alex Smola

Hierarchical Bayesian models often capture distributions over a very large number of distinct atoms. The need for these models arises when organizing huge amount of unsupervised data, for instance, features extracted using deep convnets that can be exploited to organize abundant unlabeled images. Inference for hierarchical Bayesian models in such cases can be rather nontrivial, leading to approximate approaches. In this work, we propose Canopy, a sampler based on Cover Trees that is exact, has guaranteed runtime logarithmic in the number of atoms, and is provably polynomial in the inherent dimensionality of the underlying parameter space. In other words, the algorithm is as fast as search over a hierarchical data structure. We provide theory for Canopy and demonstrate its effectiveness on both synthetic and real datasets, consisting of over 100 million images.

****

Continual Learning Through Synaptic Intelligence

Friedemann Zenke, Ben Poole, Surya Ganguli

While deep learning has led to remarkable advances across diverse applications, it struggles in domains where the data distribution changes over the course of learning. In stark contrast, biological neural networks continually adapt to changing domains, possibly by leveraging complex molecular machinery to solve many tasks simultaneously. In this study, we introduce intelligent synapses that bring some of this biological complexity into artificial neural networks. Each synapse accumulates task relevant information over time, and exploits this information to rapidly store new memories without forgetting old ones. We evaluate our approach on continual learning of classification tasks, and show that it dramatically reduces forgetting while maintaining computational efficiency.

****

Stochastic Gradient Monomial Gamma Sampler

Yizhe Zhang, Changyou Chen, Zhe Gan, Ricardo Henao, Lawrence Carin

Scaling Markov Chain Monte Carlo (MCMC) to estimate posterior distributions from large datasets has been made possible as a result of advances in stochastic gradient techniques. Despite their success, mixing performance of existing methods when sampling from multimodal distributions can be less efficient with insufficient Monte Carlo samples; this is evidenced by slow convergence and insufficient exploration of posterior distributions. We propose a generalized framework to improve the sampling efficiency of stochastic gradient MCMC, by leveraging a gener alized kinetics that delivers superior stationary mixing, especially in multimodal distributions, and propose several techniques to overcome the practical issues. We show that the proposed approach is better at exploring a complicated multimodal posterior distribution, and demonstrate improvements over other stochastic gradient MCMC methods on various applications.

****

Adversarial Feature Matching for Text Generation

Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, Lawrence Carin

The Generative Adversarial Network (GAN) has achieved great success in generating realistic (real-valued) synthetic data. However, convergence issues and difficulties dealing with discrete data hinder the applicability of GAN to text. We propose a framework for generating realistic text via adversarial training. We employ a long short-term memory network as generator, and a convolutional network as discriminator. Instead of using the standard objective of GAN, we propose matching the high-dimensional latent feature distributions of real and synthetic sentences, via a kernelized discrepancy metric. This eases adversarial training by alleviating the mode-collapsing problem. Our experiments show superior performance in quantitative evaluation, and demonstrate that our model can generate realistic-looking sentences.

****

Scaling Up Sparse Support Vector Machines by Simultaneous Feature and Sample Reduction

Weizhong Zhang, Bin Hong, Wei Liu, Jieping Ye, Deng Cai, Xiaofei He, Jie Wang

Sparse support vector machine (SVM) is a popular classification technique that can simultaneously learn a small set of the most interpretable features and identify the support vectors. It has achieved great successes in many real-world applications. However, for large-scale problems involving a huge number of samples and extremely high-dimensional features, solving sparse SVMs remains challenging. By noting that sparse SVMs induce sparsities in both feature and sample spaces, we propose a novel approach, which is based on accurate estimations of the primal and dual optima of sparse SVMs, to simultaneously identify the features and samples that are guaranteed to be irrelevant to the outputs. Thus, we can remove the identified inactive samples and features from the training phase, leading to substantial savings in both the memory usage and computational cost without sacrificing accuracy. To the best of our knowledge, the proposed method is the first static feature and sample reduction method for sparse SVMs. Experiments on both synthetic and real datasets (e.g., the kddb dataset with about 20 million samples and 30 million features) demonstrate that our approach significantly outperforms state-of-the-art methods and the speedup gained by our approach can be orders of magnitude.

****

Re-revisiting Learning on Hypergraphs: Confidence Interval and Subgradient Method

Chenzi Zhang, Shuguang Hu, Zhihao Gavin Tang, T-H. Hubert Chan

We revisit semi-supervised learning on hypergraphs. Same as previous approaches, our method uses a convex program whose objective function is not everywhere differentiable. We exploit the non-uniqueness of the optimal solutions, and consider confidence intervals which give the exact ranges that unlabeled vertices take in any optimal solution. Moreover, we give a much simpler approach for solving the convex program based on the subgradient method. Our experiments on real-world datasets confirm that our confidence interval approach on hypergraphs outperforms existing methods, and our sub-gradient method gives faster running times when the number of vertices is much larger than the number of edges.

****

ZipML: Training Linear Models with End-to-End Low Precision, and a Little Bit of Deep Learning

Hantian Zhang, Jerry Li, Kaan Kara, Dan Alistarh, Ji Liu, Ce Zhang

Recently there has been significant interest in training machine-learning models at low precision: by reducing precision, one can reduce computation and communi cation by one order of magnitude. We examine training at reduced precision, both from a theoretical and practical perspective, and ask: is it possible to train models at end-to-end low precision with provable guarantees? Can this lead to consistent order-of-magnitude speedups? We mainly focus on linear models, and the answer is yes for linear models. We develop a simple framework called ZipML based on one simple but novel strategy called double sampling. Our ZipML framework is able to execute training at low precision with no bias, guaranteeing convergence, whereas naive quantization would introduce significant bias. We validate our framework across a range of applications, and show that it enables an FPGA prototype that is up to $6.5\times$ faster than an implementation using full 32-bit precision. We further develop a variance-optimal stochastic quantization strategy and show that it can make a significant difference in a variety of settings. When applied to linear models together with double sampling, we save up to another $1.7\times$ in data movement compared with uniform quantization. When training deep networks with quantized models, we achieve higher accuracy than the state-of-the-art XNOR-Net.

****

Convexified Convolutional Neural Networks

Yuchen Zhang, Percy Liang, Martin J. Wainwright

We describe the class of convexified convolutional neural networks (CCNNs), which capture the parameter sharing of convolutional neural networks in a convex manner. By representing the nonlinear convolutional filters as vectors in a reproducing kernel Hilbert space, the CNN parameters can be represented as a low-rank matrix, which can be relaxed to obtain a convex optimization problem. For learning two-layer convolutional neural networks, we prove that the generalization error obtained by a convexified CNN converges to that of the best possible CNN. For learning deeper networks, we train CCNNs in a layer-wise manner. Empirically, CCNNs achieve competitive or better performance than CNNs trained by backpropagation, SVMs, fully-connected neural networks, stacked denoising auto-encoders, and other baseline methods.

****

Projection-free Distributed Online Learning in Networks

Wenpeng Zhang, Peilin Zhao, Wenwu Zhu, Steven C. H. Hoi, Tong Zhang

The conditional gradient algorithm has regained a surge of research interest in recent years due to its high efficiency in handling large-scale machine learning problems. However, none of existing studies has explored it in the distributed online learning setting, where locally light computation is assumed. In this paper, we fill this gap by proposing the distributed online conditional gradient algorithm, which eschews the expensive projection operation needed in its counterpart algorithms by exploiting much simpler linear optimization steps. We give a regret bound for the proposed algorithm as a function of the network size and topology, which will be smaller on smaller graphs or “well-connected” graphs. Experiments on two large-scale real-world datasets for a multiclass classification task confirm the computational benefit of the proposed algorithm and also verify the theoretical regret bound.

****

Multi-Class Optimal Margin Distribution Machine

Teng Zhang, Zhi-Hua Zhou

Recent studies disclose that maximizing the minimum margin like support vector machines does not necessarily lead to better generalization performances, and instead, it is crucial to optimize the margin distribution. Although it has been shown that for binary classification, characterizing the margin distribution by the first- and second-order statistics can achieve superior performance. It still remains open for multi-class classification, and due to the complexity of margin for multi-class classification, optimizing its distribution by mean and variance can also be difficult. In this paper, we propose mcODM (multi-class Optimal margin Distribution Machine), which can solve this problem efficiently. We also give a theoretical analysis for our method, which verifies the significance of margin distribution for multi-class classification. Empirical study further shows that mcODM always outperforms all four versions of multi-class SVMs on all experi mental data sets.

****

Leveraging Node Attributes for Incomplete Relational Data

He Zhao, Lan Du, Wray Buntine

Relational data are usually highly incomplete in practice, which inspires us to leverage side information to improve the performance of community detection and link prediction. This paper presents a Bayesian probabilistic approach that incorporates various kinds of node attributes encoded in binary form in relational models with Poisson likelihood. Our method works flexibly with both directed and undirected relational networks. The inference can be done by efficient Gibbs sampling which leverages sparsity of both networks and node attributes. Extensive experiments show that our models achieve the state-of-the-art link prediction results, especially with highly incomplete relational data.

****

Theoretical Properties for Neural Networks with Weight Matrices of Low Displacement Rank

Liang Zhao, Siyu Liao, Yanzhi Wang, Zhe Li, Jian Tang, Bo Yuan

Recently low displacement rank (LDR) matrices, or so-called structured matrices, have been proposed to compress large-scale neural networks. Empirical results have shown that neural networks with weight matrices of LDR matrices, referred as LDR neural networks, can achieve significant reduction in space and computational complexity while retaining high accuracy. This paper gives theoretical study on LDR neural networks. First, we prove the universal approximation property of LDR neural networks with a mild condition on the displacement operators. We then show that the error bounds of LDR neural networks are as efficient as general neural networks with both single-layer and multiple-layer structure. Finally, we propose back-propagation based training algorithm for general LDR neural networks.

****

Learning Hierarchical Features from Deep Generative Models

Shengjia Zhao, Jiaming Song, Stefano Ermon

Deep neural networks have been shown to be very successful at learning feature hierarchies in supervised learning tasks. Generative models, on the other hand, have benefited less from hierarchical models with multiple layers of latent variables. In this paper, we prove that hierarchical latent variable models do not take advantage of the hierarchical structure when trained with existing variational methods, and provide some limitations on the kind of features existing models can learn. Finally we propose an alternative architecture that do not suffer from these limitations. Our model is able to learn highly interpretable and disentangled hierarchical features on several natural image datasets with no task specific regularization or prior knowledge.

****

Learning Sleep Stages from Radio Signals: A Conditional Adversarial Architecture

Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S. Jaakkola, Matt T. Bianchi

We focus on predicting sleep stages from radio measurements without any attached sensors on subjects. We introduce a new predictive model that combines convolutional and recurrent neural networks to extract sleep-specific subject-invariant features from RF signals and capture the temporal progression of sleep. A key innovation underlying our approach is a modified adversarial training regime that discards extraneous information specific to individuals or measurement conditions, while retaining all information relevant to the predictive task. We analyze our game theoretic setup and empirically demonstrate that our model achieves significant improvements over state-of-the-art solutions.

****

Follow the Moving Leader in Deep Learning

Shuai Zheng, James T. Kwok

Deep networks are highly nonlinear and difficult to optimize. During training, the parameter iterate may move from one local basin to another, or the data distribution may even change. Inspired by the close connection between stochastic optimization and online learning, we propose a variant of the follow the regularize d leader (FTRL) algorithm called follow the moving leader (FTML). Unlike the FTRL family of algorithms, the recent samples are weighted more heavily in each iteration and so FTML can adapt more quickly to changes. We show that FTML enjoys the nice properties of RMSprop and Adam, while avoiding their pitfalls. Experimental results on a number of deep learning models and tasks demonstrate that FTML converges quickly, and outperforms other state-of-the-art optimizers.

****

Asynchronous Stochastic Gradient Descent with Delay Compensation

Shuxin Zheng, Qi Meng, Taifeng Wang, Wei Chen, Nenghai Yu, Zhi-Ming Ma, Tie-Yan Liu

With the fast development of deep learning, it has become common to learn big neural networks using massive training data. Asynchronous Stochastic Gradient Descent (ASGD) is widely adopted to fulfill this task for its efficiency, which is, however, known to suffer from the problem of delayed gradients. That is, when a local worker adds its gradient to the global model, the global model may have been updated by other workers and this gradient becomes “delayed”. We propose a novel technology to compensate this delay, so as to make the optimization behavior of ASGD closer to that of sequential SGD. This is achieved by leveraging Taylor expansion of the gradient function and efficient approximators to the Hessian matrix of the loss function. We call the new algorithm Delay Compensated ASGD (DC-ASGD). We evaluated the proposed algorithm on CIFAR-10 and ImageNet datasets, and the experimental results demonstrate that DC-ASGD outperforms both synchronous SGD and asynchronous SGD, and nearly approaches the performance of sequential SGD.

****

Collect at Once, Use Effectively: Making Non-interactive Locally Private Learning Possible

Kai Zheng, Wenlong Mou, Liwei Wang

Non-interactive Local Differential Privacy (LDP) requires data analysts to collect data from users through noisy channel at once. In this paper, we extend the frontiers of Non-interactive LDP learning and estimation from several aspects. For learning with smooth generalized linear losses, we propose an approximate stochastic gradient oracle estimated from non-interactive LDP channel using Chebyshev expansion, which is combined with inexact gradient methods to obtain an efficient algorithm with quasi-polynomial sample complexity bound. For the high-dimensional world, we discover that under ℓ_2-norm assumption on data points, high-dimensional sparse linear regression and mean estimation can be achieved with logarithmic dependence on dimension, using random projection and approximate recovery. We also extend our methods to Kernel Ridge Regression. Our work is the first one that makes learning and estimation possible for a broad range of learning tasks under non-interactive LDP model.

****

Recovery Guarantees for One-hidden-layer Neural Networks

Kai Zhong, Zhao Song, Prateek Jain, Peter L. Bartlett, Inderjit S. Dhillon

In this paper, we consider regression problems with one-hidden-layer neural networks (1NNs). We distill some properties of activation functions that lead to local strong convexity in the neighborhood of the ground-truth parameters for the 1NN squared-loss objective and most popular nonlinear activation functions satisfy the distilled properties, including rectified linear units (ReLUs), leaky ReLUs, squared ReLUs and sigmoids. For activation functions that are also smooth, we show local linear convergence guarantees of gradient descent under a resampling rule. For homogeneous activations, we show tensor methods are able to initialize the parameters to fall into the local strong convexity region. As a result, tensor initialization followed by gradient descent is guaranteed to recover the ground truth with sample complexity $ d \cdot \log(1/\epsilon) \cdot \mathrm{poly}(k,\lambda)$ and computational complexity $n\cdot d \cdot \mathrm{poly}(k,\lambda) $ for smooth homogeneous activations with high probability, where d is the dimension of the input, k ($k\leq d$) is the number of hidden nodes, λ is a conditioning property of the ground-truth parameter matrix between the input layer and the hidden layer, ϵ is the targeted precision and $n

$ is the number of samples. To the best of our knowledge, this is the first work that provides recovery guarantees for 1NNs with both sample complexity and computational complexity linear in the input dimension and logarithmic in the precision.

****

Stochastic Adaptive Quasi-Newton Methods for Minimizing Expected Values

Chaoxu Zhou, Wenbo Gao, Donald Goldfarb

We propose a novel class of stochastic, adaptive methods for minimizing self-concordant functions which can be expressed as an expected value. These methods generate an estimate of the true objective function by taking the empirical mean over a sample drawn at each step, making the problem tractable. The use of adaptive step sizes eliminates the need for the user to supply a step size. Methods in this class include extensions of gradient descent (GD) and BFGS. We show that, given a suitable amount of sampling, the stochastic adaptive GD attains linear convergence in expectation, and with further sampling, the stochastic adaptive BFGS attains R-superlinear convergence. We present experiments showing that these methods compare favorably to SGD.

****

Identify the Nash Equilibrium in Static Games with Random Payoffs

Yichi Zhou, Jialian Li, Jun Zhu

We study the problem on how to learn the pure Nash Equilibrium of a two-player zero-sum static game with random payoffs under unknown distributions via efficient payoff queries. We introduce a multi-armed bandit model to this problem due to its ability to find the best arm efficiently among random arms and propose two algorithms for this problem—LUCB-G based on the confidence bounds and a racing algorithm based on successive action elimination. We provide an analysis on the sample complexity lower bound when the Nash Equilibrium exists.

****

When can Multi-Site Datasets be Pooled for Regression? Hypothesis Tests, ℓ_2-consistency and Neuroscience Applications

Hao Henry Zhou, Yilin Zhang, Vamsi K. Ithapu, Sterling C. Johnson, Grace Wahba, Vikas Singh

Many studies in biomedical and health sciences involve small sample sizes due to logistic or financial constraints. Often, identifying weak (but scientifically interesting) associations between a set of predictors and a response necessitates pooling datasets from multiple diverse labs or groups. While there is a rich literature in statistical machine learning to address distributional shifts and inference in multi-site datasets, it is less clear when such pooling is guaranteed to help (and when it does not) – independent of the inference algorithms we use. In this paper, we present a hypothesis test to answer this question, both for classical and high dimensional linear regression. We precisely identify regimes where pooling datasets across multiple sites is sensible, and how such policy decisions can be made via simple checks executable on each site before any data transfer ever happens. With a focus on Alzheimer’s disease studies, we present empirical results showing that in regimes suggested by our analysis, pooling a local dataset with data from an international study improves power.

****

High-Dimensional Variance-Reduced Stochastic Gradient Expectation-Maximization Algorithm

Rongda Zhu, Lingxiao Wang, Chengxiang Zhai, Quanquan Gu

We propose a generic stochastic expectation-maximization (EM) algorithm for the estimation of high-dimensional latent variable models. At the core of our algorithm is a novel semi-stochastic variance-reduced gradient designed for the Q-function in the EM algorithm. Under a mild condition on the initialization, our algorithm is guaranteed to attain a linear convergence rate to the unknown parameter of the latent variable model, and achieve an optimal statistical rate up to a logarithmic factor for parameter estimation. Compared with existing high-dimensional EM algorithms, our algorithm enjoys a better computational complexity and is therefore more efficient. We apply our generic algorithm to two illustrative latent variable models: Gaussian mixture model and mixture of linear regression, and demonstrate the advantages of our algorithm by both theoretical analysis and numerical experiments. We believe that the proposed semi-stochastic gradient is of independent interest for general nonconvex optimization problems with bivariate structures.

****

Recurrent Highway Networks

Julian Georg Zilly, Rupesh Kumar Srivastava, Jan Koutn■■k, Jürgen Schmidhuber

Many sequential processing tasks require complex nonlinear transition functions from one step to the next. However, recurrent neural networks with “deep” transition functions remain difficult to train, even when using Long Short-Term Memory (LSTM) networks. We introduce a novel theoretical analysis of recurrent networks based on Gersgorin’s circle theorem that illuminates several modeling and optimization issues and improves our understanding of the LSTM cell. Based on this analysis we propose Recurrent Highway Networks, which extend the LSTM architecture to allow step-to-step transition depths larger than one. Several language modeling experiments demonstrate that the proposed architecture results in powerful and efficient models. On the Penn Treebank corpus, solely increasing the transition depth from 1 to 10 improves word-level perplexity from 90.6 to 65.4 using the same number of parameters. On the larger Wikipedia datasets for character prediction (text8 and enwik8), RHNs outperform all previous results and achieve an entropy of 1.27 bits per character.

****

Online Learning to Rank in Stochastic Click Models

Masrour Zoghi, Tomas Tunys, Mohammad Ghavamzadeh, Branislav Kveton, Csaba Szepesvari, Zheng Wen

Online learning to rank is a core problem in information retrieval and machine learning. Many provably efficient algorithms have been recently proposed for this problem in specific click models. The click model is a model of how the user interacts with a list of documents. Though these results are significant, their impact on practice is limited, because all proposed algorithms are designed for specific click models and lack convergence guarantees in other models. In this work, we propose BatchRank, the first online learning to rank algorithm for a broad class of click models. The class encompasses two most fundamental click models, the cascade and position-based models. We derive a gap-dependent upper bound on the T-step regret of BatchRank and evaluate it on a range of web search queries. We observe that BatchRank outperforms ranked bandits and is more robust than CascadeKL-UCB, an existing algorithm for the cascade model.

****