Sign in to GitHub · GitHub
****
No Oops, You Won’t Do It Again: Mechanisms for Self-correction in Crowdsourcing
Nihar Shah, Dengyong Zhou
Crowdsourcing is a very popular means of obtaining the large amounts of labeled data that modern machine learning methods require. Although cheap and fast to obtain, crowdsourced labels suffer from significant amounts of error, thereby degrading the performance of downstream machine learning tasks. With the goal of improving the quality of the labeled data, we seek to mitigate the many errors that occur due to silly mistakes or inadvertent errors by crowdsourcing workers. We propose a two-stage setting for crowdsourcing where the worker first answers the questions, and is then allowed to change her answers after looking at a (noisy) reference answer. We mathematically formulate this process and develop mechanisms to incentivize workers to act appropriately. Our mathematical guarantees show that our mechanism incentivizes the workers to answer honestly in both stages, and refrain from answering randomly in the first stage or simply copying in the second. Numerical experiments reveal a significant boost in performance that such "self-correction" can provide when using crowdsourcing to train machine learning algorithms.
****
Stochastically Transitive Models for Pairwise Comparisons: Statistical and Computational Issues
Nihar Shah, Sivaraman Balakrishnan, Aditya Guntuboyina, Martin Wainwright
There are various parametric models for analyzing pairwise comparison data, including the Bradley-Terry-Luce (BTL) and Thurstone models, but their reliance on strong parametric assumptions is limiting. In this work, we study a flexible model for pairwise comparisons, under which the probabilities of outcomes are required only to satisfy a natural form of stochastic transitivity. This class includes parametric models including the BTL and Thurstone models as special cases, but is considerably more general. We provide various examples of models in this broader stochastically transitive class for which classical parametric models provide poor fits. Despite this greater flexibility, we show that the matrix of probabilities can be estimated at the same rate as in standard parametric models. On the other hand, unlike in the BTL and Thurstone models, computing the minimax-optimal estimator in the stochastically transitive model is non-trivial, and we explore various computationally tractable alternatives. We show that a simple singular value thresholding algorithm is statistically consistent but does not achieve the minimax rate. We then propose and study algorithms that achieve the minimax rate over interesting sub-classes of the full stochastically transitive class. We complement our theoretical results with thorough numerical simulations.
****
Uprooting and Rerooting Graphical Models
Adrian Weller
We show how any binary pairwise model may be “uprooted” to a fully symmetric model, wherein original singleton potentials are transformed to potentials on edges to an added variable, and then “rerooted” to a new model on the original number of variables. The new model is essentially equivalent to the original model, with the same partition function and allowing recovery of the original marginals or a MAP configuration, yet may have very different computational properties that allow much more efficient inference. This meta-approach deepens our understanding, may be applied to any existing algorithm to yield improved methods in practice, generalizes earlier theoretical results, and reveals a remarkable interpretation of the triplet-consistent polytope.
****
A Deep Learning Approach to Unsupervised Ensemble Learning
Uri Shaham, Xiuyuan Cheng, Omer Dror, Ariel Jaffe, Boaz Nadler, Joseph Chang, Yuval Kluger
We show how deep learning methods can be applied in the context of crowdsourcing and unsupervised ensemble learning. First, we prove that the popular model of Dawid and Skene, which assumes that all classifiers are conditionally independent

, is \em equivalent to a Restricted Boltzmann Machine (RBM) with a single hidden node. Hence, under this model, the posterior probabilities of the true labels can be instead estimated via a trained RBM. Next, to address the more general case, where classifiers may strongly violate the conditional independence assumption, we propose to apply RBM-based Deep Neural Net (DNN). Experimental results on various simulated and real-world datasets demonstrate that our proposed DNN approach outperforms other state-of-the-art methods, in particular when the data violates the conditional independence assumption.
****
Revisiting Semi-Supervised Learning with Graph Embeddings
Zhilin Yang, William Cohen, Ruslan Salakhudinov
We present a semi-supervised learning framework based on graph embeddings. Given a graph between instances, we train an embedding for each instance to jointly predict the class label and the neighborhood context in the graph. We develop both transductive and inductive variants of our method. In the transductive variant of our method, the class labels are determined by both the learned embeddings and input feature vectors, while in the inductive variant, the embeddings are defined as a parametric function of the feature vectors, so predictions can be made on instances not seen during training. On a large and diverse set of benchmark tasks, including text classification, distantly supervised entity extraction, and entity classification, we show improved performance over many of the existing models.
****
Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization
Chelsea Finn, Sergey Levine, Pieter Abbeel
Reinforcement learning can acquire complex behaviors from high-level specifications. However, defining a cost function that can be optimized effectively and encodes the correct task is challenging in practice. We explore how inverse optimal control (IOC) can be used to learn behaviors from demonstrations, with applications to torque control of high-dimensional robotic systems. Our method addresses two key challenges in inverse optimal control: first, the need for informative features and effective regularization to impose structure on the cost, and second, the difficulty of learning the cost function under unknown dynamics for high-dimensional continuous systems. To address the former challenge, we present an algorithm capable of learning arbitrary nonlinear cost functions, such as neural networks, without meticulous feature engineering. To address the latter challenge, we formulate an efficient sample-based approximation for MaxEnt IOC. We evaluate our method on a series of simulated tasks and real-world robotic manipulation problems, demonstrating substantial improvement over prior methods both in terms of task complexity and sample efficiency.
****
Diversity-Promoting Bayesian Learning of Latent Variable Models
Pengtao Xie, Jun Zhu, Eric Xing
In learning latent variable models (LVMs), it is important to effectively capture infrequent patterns and shrink model size without sacrificing modeling power. Various studies have been done to “diversify” a LVM, which aim to learn a diverse set of latent components in LVMs. Most existing studies fall into a frequentist-style regularization framework, where the components are learned via point estimation. In this paper, we investigate how to “diversify” LVMs in the paradigm of Bayesian learning, which has advantages complementary to point estimation, such as alleviating overfitting via model averaging and quantifying uncertainty. We propose two approaches that have complementary advantages. One is to define diversity-promoting mutual angular priors which assign larger density to components with larger mutual angles based on Bayesian network and von Mises-Fisher distribution and use these priors to affect the posterior via Bayes rule. We develop two efficient approximate posterior inference algorithms based on variational inference and Markov chain Monte Carlo sampling. The other approach is to impose diversity-promoting regularization directly over the post-data distribution of components. These two methods are applied to the Bayesian mixture of experts model to encourage the “experts” to be diverse and experimental results demonstrate th

e effectiveness and efficiency of our methods.
****
Additive Approximations in High Dimensional Nonparametric Regression via the SALSA
Kirthevasan Kandasamy, Yaoliang Yu
High dimensional nonparametric regression is an inherently difficult problem with known lower bounds depending exponentially in dimension. A popular strategy to alleviate this curse of dimensionality has been to use additive models of \emph first order, which model the regression function as a sum of independent functions on each dimension. Though useful in controlling the variance of the estimate, such models are often too restrictive in practical settings. Between non-additive models which often have large variance and first order additive models which have large bias, there has been little work to exploit the trade-off in the middle via additive models of intermediate order. In this work, we propose salsa, which bridges this gap by allowing interactions between variables, but controls model capacity by limiting the order of interactions. salsas minimises the residual sum of squares with squared RKHS norm penalties. Algorithmically, it can be viewed as Kernel Ridge Regression with an additive kernel. When the regression function is additive, the excess risk is only polynomial in dimension. Using the Girard-Newton formulae, we efficiently sum over a combinatorial number of terms in the additive expansion. Via a comparison on 15 real datasets, we show that our method is competitive against 21 other alternatives.
****
Hawkes Processes with Stochastic Excitations
Young Lee, Kar Wai Lim, Cheng Soon Ong
We propose an extension to Hawkes processes by treating the levels of self-excitation as a stochastic differential equation. Our new point process allows better approximation in application domains where events and intensities accelerate each other with correlated levels of contagion. We generalize a recent algorithm for simulating draws from Hawkes processes whose levels of excitation are stochastic processes, and propose a hybrid Markov chain Monte Carlo approach for model fitting. Our sampling procedure scales linearly with the number of required events and does not require stationarity of the point process. A modular inference procedure consisting of a combination between Gibbs and Metropolis Hastings steps is put forward. We recover expectation maximization as a special case. Our general approach is illustrated for contagion following geometric Brownian motion and exponential Langevin dynamics.
****
Data-driven Rank Breaking for Efficient Rank Aggregation
Ashish Khetan, Sewoong Oh
Rank aggregation systems collect ordinal preferences from individuals to produce a global ranking that represents the social preference. To reduce the computational complexity of learning the global ranking, a common practice is to use rank-breaking. Individuals’ preferences are broken into pairwise comparisons and then applied to efficient algorithms tailored for independent pairwise comparisons. However, due to the ignored dependencies, naive rank-breaking approaches can result in inconsistent estimates. The key idea to produce unbiased and accurate estimates is to treat the paired comparisons outcomes unequally, depending on the topology of the collected data. In this paper, we provide the optimal rank-breaking estimator, which not only achieves consistency but also achieves the best error bound. This allows us to characterize the fundamental tradeoff between accuracy and complexity in some canonical scenarios. Further, we identify how the accuracy depends on the spectral gap of a corresponding comparison graph.
****
Dropout distillation
Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder
Dropout is a popular stochastic regularization technique for deep neural networks that works by randomly dropping (i.e. zeroing) units from the network during training. This randomization process allows to implicitly train an ensemble of exponentially many networks sharing the same parametrization, which should be aver

aged at test time to deliver the final prediction. A typical workaround for this intractable averaging operation consists in scaling the layers undergoing dropout randomization. This simple rule called ’standard dropout’ is efficient, but might degrade the accuracy of the prediction. In this work we introduce a novel approach, coined ’dropout distillation’, that allows us to train a predictor in a way to better approximate the intractable, but preferable, averaging process, while keeping under control its computational efficiency. We are thus able to construct models that are as efficient as standard dropout, or even more efficient, while being more accurate. Experiments on standard benchmark datasets demonstrate the validity of our method, yielding consistent improvements over conventional dropout.

****

Metadata-conscious anonymous messaging

Giulia Fanti, Peter Kairouz, Sewoong Oh, Kannan Ramchandran, Pramod Viswanath

Anonymous messaging platforms like Whisper and Yik Yak allow users to spread messages over a network (e.g., a social network) without revealing message authorship to other users. The spread of messages on these platforms can be modeled by a diffusion process over a graph. Recent advances in network analysis have revealed that such diffusion processes are vulnerable to author deanonymization by adversaries with access to metadata, such as timing information. In this work, we ask the fundamental question of how to propagate anonymous messages over a graph to make it difficult for adversaries to infer the source. In particular, we study the performance of a message propagation protocol called adaptive diffusion introduced in (Fanti et al., 2015). We prove that when the adversary has access to metadata at a fraction of corrupted graph nodes, adaptive diffusion achieves asymptotically optimal source-hiding and significantly outperforms standard diffusion. We further demonstrate empirically that adaptive diffusion hides the source effectively on real social networks.

****

The Teaching Dimension of Linear Learners

Ji Liu, Xiaojin Zhu, Hrag Ohannessian

Teaching dimension is a learning theoretic quantity that specifies the minimum training set size to teach a target model to a learner. Previous studies on teaching dimension focused on version-space learners which maintain all hypotheses consistent with the training data, and cannot be applied to modern machine learners which select a specific hypothesis via optimization. This paper presents the first known teaching dimension for ridge regression, support vector machines, and logistic regression. We also exhibit optimal training sets that match these teaching dimensions. Our approach generalizes to other linear learners.

****

Truthful Univariate Estimators

Ioannis Caragiannis, Ariel Procaccia, Nisarg Shah

We revisit the classic problem of estimating the population mean of an unknown single-dimensional distribution from samples, taking a game-theoretic viewpoint. In our setting, samples are supplied by strategic agents, who wish to pull the estimate as close as possible to their own value. In this setting, the sample mean gives rise to manipulation opportunities, whereas the sample median does not. Our key question is whether the sample median is the best (in terms of mean squared error) truthful estimator of the population mean. We show that when the underlying distribution is symmetric, there are truthful estimators that dominate the median. Our main result is a characterization of worst-case optimal truthful estimators, which provably outperform the median, for possibly asymmetric distributions with bounded support.

****

Why Regularized Auto-Encoders learn Sparse Representation?

Devansh Arpit, Yingbo Zhou, Hung Ngo, Venu Govindaraju

Sparse distributed representation is the key to learning useful features in deep learning algorithms, because not only it is an efficient mode of data representation, but also – more importantly – it captures the generation process of most real world data. While a number of regularized auto-encoders (AE) enforce sparsi

ty explicitly in their learned representation and others don’t, there has been l
ittle formal analysis on what encourages sparsity in these models in general. Ou
r objective is to formally study this general problem for regularized auto-encod
ers. We provide sufficient conditions on both regularization and activation func
tions that encourage sparsity. We show that multiple popular models (de-noising
 and contractive auto encoders, e.g.) and activations (rectified linear and sigmo
id, e.g.) satisfy these conditions; thus, our conditions help explain sparsity i
n their learned representation. Thus our theoretical and empirical analysis toge
ther shed light on the properties of regularization/activation that are conducti
ve to sparsity and unify a number of existing auto-encoder models and activation
 functions under the same analytical framework.
****

k-variates++: more pluses in the k-means++
Richard Nock, Raphael Canyasse, Roksana Boreli, Frank Nielsen
k-means++ seeding has become a de facto standard for hard clustering algorithms.
 In this paper, our first contribution is a two-way generalisation of this seedi
ng, k-variates++, that includes the sampling of general densities rather than ju
st a discrete set of Dirac densities anchored at the point locations, *and* a ge
neralisation of the well known Arthur-Vassilvitskii (AV) approximation guarantee
, in the form of a *bias+variance* approximation bound of the *global* optimum.
This approximation exhibits a reduced dependency on the "noise" component with r
espect to the optimal potential — actually approaching the statistical lower bou
nd. We show that k-variates++ *reduces* to efficient (biased seeding) clustering
 algorithms tailored to specific frameworks; these include distributed, streamin
g and on-line clustering, with *direct* approximation results for these algorith
ms. Finally, we present a novel application of k-variates++ to differential priv
acy. For either the specific frameworks considered here, or for the differential
 privacy setting, there is little to no prior results on the direct application
of k-means++ and its approximation bounds — state of the art contenders appear t
o be significantly more complex and / or display less favorable (approximation)
properties. We stress that our algorithms can still be run in cases where there
is *no* closed form solution for the population minimizer. We demonstrate the ap
plicability of our analysis via experimental evaluation on several domains and s
ettings, displaying competitive performances vs state of the art.
****

Multi-Player Bandits – a Musical Chairs Approach
Jonathan Rosenski, Ohad Shamir, Liran Szlak
We consider a variant of the stochastic multi-armed bandit problem, where multip
le players simultaneously choose from the same set of arms and may collide, rece
iving no reward. This setting has been motivated by problems arising in cognitiv
e radio networks, and is especially challenging under the realistic assumption t
hat communication between players is limited. We provide a communication-free al
gorithm (Musical Chairs) which attains constant regret with high probability, as
 well as a sublinear-regret, communication-free algorithm (Dynamic Musical Chair
s) for the more difficult setting of players dynamically entering and leaving th
roughout the game. Moreover, both algorithms do not require prior knowledge of t
he number of players. To the best of our knowledge, these are the first communic
ation-free algorithms with these types of formal guarantees.
****

The Information Sieve
Greg Ver Steeg, Aram Galstyan
We introduce a new framework for unsupervised learning of representations based
on a novel hierarchical decomposition of information. Intuitively, data is passe
d through a series of progressively fine-grained sieves. Each layer of the sieve
 recovers a single latent factor that is maximally informative about multivariat
e dependence in the data. The data is transformed after each pass so that the re
maining unexplained information trickles down to the next layer. Ultimately, we
are left with a set of latent factors explaining all the dependence in the origi
nal data and remainder information consisting of independent noise. We present a
 practical implementation of this framework for discrete variables and apply it

to a variety of fundamental tasks in unsupervised learning including independent component analysis, lossy and lossless compression, and predicting missing values in data.
****

Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin
Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, Zhenyao Zhu
We show that an end-to-end deep learning approach can be used to recognize either English or Mandarin Chinese speech–two vastly different languages. Because it replaces entire pipelines of hand-engineered components with neural networks, end-to-end learning allows us to handle a diverse variety of speech including noisy environments, accents and different languages. Key to our approach is our application of HPC techniques, enabling experiments that previously took weeks to now run in days. This allows us to iterate more quickly to identify superior architectures and algorithms. As a result, in several cases, our system is competitive with the transcription of human workers when benchmarked on standard datasets. Finally, using a technique called Batch Dispatch with GPUs in the data center, we show that our system can be inexpensively deployed in an online setting, delivering low latency when serving users at scale.
****

On the Consistency of Feature Selection With Lasso for Non-linear Targets
Yue Zhang, Weihong Guo, Soumya Ray
An important question in feature selection is whether a selection strategy recovers the “true” set of features, given enough data. We study this question in the context of the popular Least Absolute Shrinkage and Selection Operator (Lasso) feature selection strategy. In particular, we consider the scenario when the model is misspecified so that the learned model is linear while the underlying real target is nonlinear. Surprisingly, we prove that under certain conditions, Lasso is still able to recover the correct features in this case. We also carry out numerical studies to empirically verify the theoretical results and explore the necessity of the conditions under which the proof holds.
****

Minimum Regret Search for Single- and Multi-Task Optimization
Jan Hendrik Metzen
We propose minimum regret search (MRS), a novel acquisition function for Bayesian optimization. MRS bears similarities with information-theoretic approaches such as entropy search (ES). However, while ES aims in each query at maximizing the information gain with respect to the global maximum, MRS aims at minimizing the expected simple regret of its ultimate recommendation for the optimum. While empirically ES and MRS perform similar in most of the cases, MRS produces fewer outliers with high simple regret than ES. We provide empirical results both for a synthetic single-task optimization problem as well as for a simulated multi-task robotic control problem.
****

CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy
Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, John Wernsing
Applying machine learning to a problem which involves medical, financial, or other types of sensitive data, not only requires accurate predictions but also care

ful attention to maintaining data privacy and security. Legal and ethical requir ements may prevent the use of cloud-based machine learning solutions for such ta sks. In this work, we will present a method to convert learned neural networks t o CryptoNets, neural networks that can be applied to encrypted data. This allows a data owner to send their data in an encrypted form to a cloud service that ho sts the network. The encryption ensures that the data remains confidential since the cloud does not have access to the keys needed to decrypt it. Nevertheless, we will show that the cloud service is capable of applying the neural network to the encrypted data to make encrypted predictions, and also return them in encry pted form. These encrypted predictions can be sent back to the owner of the secr et key who can decrypt them. Therefore, the cloud service does not gain any info rmation about the raw data nor about the prediction it made. We demonstrate Cryp toNets on the MNIST optical character recognition tasks. CryptoNets achieve 99% accuracy and can make around 59000 predictions per hour on a single PC. Therefor e, they allow high throughput, accurate, and private predictions.
****

The Variational Nystrom method for large-scale spectral problems
Max Vladymyrov, Miguel Carreira-Perpinan
Spectral methods for dimensionality reduction and clustering require solving an eigenproblem defined by a sparse affinity matrix. When this matrix is large, one seeks an approximate solution. The standard way to do this is the Nystrom metho d, which first solves a small eigenproblem considering only a subset of landmark points, and then applies an out-of-sample formula to extrapolate the solution t o the entire dataset. We show that by constraining the original problem to satis fy the Nystrom formula, we obtain an approximation that is computationally simpl e and efficient, but achieves a lower approximation error using fewer landmarks and less runtime. We also study the role of normalization in the computational c ost and quality of the resulting solution.
****

Multi-Bias Non-linear Activation in Deep Neural Networks
Hongyang Li, Wanli Ouyang, Xiaogang Wang
As a widely used non-linear activation, Rectified Linear Unit (ReLU) separates n oise and signal in a feature map by learning a threshold or bias. However, we ar gue that the classification of noise and signal not only depends on the magnitud e of responses, but also the context of how the feature responses would be used to detect more abstract patterns in higher layers. In order to output multiple r esponse maps with magnitude in different ranges for a particular visual pattern, existing networks employing ReLU and its variants have to learn a large number of redundant filters. In this paper, we propose a multi-bias non-linear activati on (MBA) layer to explore the information hidden in the magnitudes of responses. It is placed after the convolution layer to decouple the responses to a convolu tion kernel into multiple maps by multi-thresholding magnitudes, thus generating more patterns in the feature space at a low computational cost. It provides gre at flexibility of selecting responses to different visual patterns in different magnitude ranges to form rich representations in higher layers. Such a simple an d yet effective scheme achieves the state-of-the-art performance on several benc hmarks.
****

Asymmetric Multi-task Learning Based on Task Relatedness and Loss
Giwoong Lee, Eunho Yang, Sung Hwang
We propose a novel multi-task learning method that can minimize the effect of ne gative transfer by allowing asymmetric transfer between the tasks based on task relatedness as well as the amount of individual task losses, which we refer to a s Asymmetric Multi-task Learning (AMTL). To tackle this problem, we couple multi ple tasks via a sparse, directed regularization graph, that enforces each task p arameter to be reconstructed as a sparse combination of other tasks, which are s elected based on the task-wise loss. We present two different algorithms to solv e this joint learning of the task predictors and the regularization graph. The f irst algorithm solves for the original learning objective using alternative opti mization, and the second algorithm solves an approximation of it using curriculu

m learning strategy, that learns one task at a time. We perform experiments on m ultiple datasets for classification and regression, on which we obtain significa nt improvements in performance over the single task learning and symmetric multi task learning baselines.
****
Accurate Robust and Efficient Error Estimation for Decision Trees
Lixin Fan
This paper illustrates a novel approach to the estimation of generalization erro r of decision tree classifiers. We set out the study of decision tree errors in the context of consistency analysis theory, which proved that the Bayes error ca n be achieved only if when the number of data samples thrown into each leaf node goes to infinity. For the more challenging and practical case where the sample size is finite or small, a novel sampling error term is introduced in this paper to cope with the small sample problem effectively and efficiently. Extensive ex perimental results show that the proposed error estimate is superior to the well known K-fold cross validation methods in terms of robustness and accuracy. More over it is orders of magnitudes more efficient than cross validation methods.
****
Fast Stochastic Algorithms for SVD and PCA: Convergence Properties and Convexity
Ohad Shamir
We study the convergence properties of the VR-PCA algorithm introduced by (Shami r, 2015) for fast computation of leading singular vectors. We prove several new results, including a formal analysis of a block version of the algorithm, and co nvergence from random initialization. We also make a few observations of indepen dent interest, such as how pre-initializing with just a single exact power itera tion can significantly improve the analysis, and what are the convexity and non- convexity properties of the underlying optimization problem.
****
Convergence of Stochastic Gradient Descent for PCA
Ohad Shamir
We consider the problem of principal component analysis (PCA) in a streaming sto chastic setting, where our goal is to find a direction of approximate maximal va riance, based on a stream of i.i.d. data points in R^d. A simple and computation ally cheap algorithm for this is stochastic gradient descent (SGD), which increm entally updates its estimate based on each new data point. However, due to the n on-convex nature of the problem, analyzing its performance has been a challenge. In particular, existing guarantees rely on a non-trivial eigengap assumption on the covariance matrix, which is intuitively unnecessary. In this paper, we prov ide (to the best of our knowledge) the first eigengap-free convergence guarantee s for SGD in the context of PCA. This also partially resolves an open problem po sed in (Hardt & Price, 2014). Moreover, under an eigengap assumption, we show th at the same techniques lead to new SGD convergence guarantees with better depend ence on the eigengap.
****
Dealbreaker: A Nonlinear Latent Variable Model for Educational Data
Andrew Lan, Tom Goldstein, Richard Baraniuk, Christoph Studer
Statistical models of student responses on assessment questions, such as those i n homeworks and exams, enable educators and computer-based personalized learning systems to gain insights into students' knowledge using machine learning. Popul ar student-response models, including the Rasch model and item response theory m odels, represent the probability of a student answering a question correctly usi ng an affine function of latent factors. While such models can accurately predic t student responses, their ability to interpret the underlying knowledge structu re (which is certainly nonlinear) is limited. In response, we develop a new, non linear latent variable model that we call the dealbreaker model, in which a stud ent's success probability is determined by their weakest concept mastery. We dev elop efficient parameter inference algorithms for this model using novel methods for nonconvex optimization. We show that the dealbreaker model achieves compara ble or better prediction performance as compared to affine models with real-worl d educational datasets. We further demonstrate that the parameters learned by th

e dealbreaker model are interpretable—they provide key insights into which concepts are critical (i.e., the “dealbreaker”) to answering a question correctly. We conclude by reporting preliminary results for a movie-rating dataset, which illustrate the broader applicability of the dealbreaker model.

****

A Kernelized Stein Discrepancy for Goodness-of-fit Tests

Qiang Liu, Jason Lee, Michael Jordan

We derive a new discrepancy statistic for measuring differences between two probability distributions based on combining Stein’s identity and the reproducing kernel Hilbert space theory. We apply our result to test how well a probabilistic model fits a set of observations, and derive a new class of powerful goodness-of-fit tests that are widely applicable for complex and high dimensional distributions, even for those with computationally intractable normalization constants. Both theoretical and empirical properties of our methods are studied thoroughly.

****

Variable Elimination in the Fourier Domain

Yexiang Xue, Stefano Ermon, Ronan Le Bras, Carla, Bart Selman

The ability to represent complex high dimensional probability distributions in a compact form is one of the key insights in the field of graphical models. Factored representations are ubiquitous in machine learning and lead to major computational advantages. We explore a different type of compact representation based on discrete Fourier representations, complementing the classical approach based on conditional independencies. We show that a large class of probabilistic graphical models have a compact Fourier representation. This theoretical result opens up an entirely new way of approximating a probability distribution. We demonstrate the significance of this approach by applying it to the variable elimination algorithm. Compared with the traditional bucket representation and other approximate inference algorithms, we obtain significant improvements.

****

Low-Rank Matrix Approximation with Stability

Dongsheng Li, Chao Chen, Qin Lv, Junchi Yan, Li Shang, Stephen Chu

Low-rank matrix approximation has been widely adopted in machine learning applications with sparse data, such as recommender systems. However, the sparsity of the data, incomplete and noisy, introduces challenges to the algorithm stability – small changes in the training data may significantly change the models. As a result, existing low-rank matrix approximation solutions yield low generalization performance, exhibiting high error variance on the training dataset, and minimizing the training error may not guarantee error reduction on the testing dataset. In this paper, we investigate the algorithm stability problem of low-rank matrix approximations. We present a new algorithm design framework, which (1) introduces new optimization objectives to guide stable matrix approximation algorithm design, and (2) solves the optimization problem to obtain stable low-rank approximation solutions with good generalization performance. Experimental results on real-world datasets demonstrate that the proposed work can achieve better prediction accuracy compared with both state-of-the-art low-rank matrix approximation methods and ensemble methods in recommendation task.

****

Linking losses for density ratio and class-probability estimation

Aditya Menon, Cheng Soon Ong

Given samples from two densities p and q, density ratio estimation (DRE) is the problem of estimating the ratio p/q. Two popular discriminative approaches to DRE are KL importance estimation (KLIEP), and least squares importance fitting (LSIF). In this paper, we show that KLIEP and LSIF both employ class-probability estimation (CPE) losses. Motivated by this, we formally relate DRE and CPE, and demonstrate the viability of using existing losses from one problem for the other. For the DRE problem, we show that essentially any CPE loss (eg logistic, exponential) can be used, as this equivalently minimises a Bregman divergence to the true density ratio. We show how different losses focus on accurately modelling different ranges of the density ratio, and use this to design new CPE losses for DRE. For the CPE problem, we argue that the LSIF loss is useful in the regime whe

re one wishes to rank instances with maximal accuracy at the head of the ranking . In the course of our analysis, we establish a Bregman divergence identity that may be of independent interest.
****

Stochastic Variance Reduction for Nonconvex Optimization
Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, Alex Smola
We study nonconvex finite-sum problems and analyze stochastic variance reduced g radient (SVRG) methods for them. SVRG and related methods have recently surged i nto prominence for convex optimization given their edge over stochastic gradient descent (SGD); but their theoretical analysis almost exclusively assumes convex ity. In contrast, we prove non-asymptotic rates of convergence (to stationary po ints) of SVRG for nonconvex optimization, and show that it is provably faster th an SGD and gradient descent. We also analyze a subclass of nonconvex problems on which SVRG attains linear convergence to the global optimum. We extend our anal ysis to mini-batch variants of SVRG, showing (theoretical) linear speedup due to minibatching in parallel settings.
****

Hierarchical Variational Models
Rajesh Ranganath, Dustin Tran, David Blei
Black box variational inference allows researchers to easily prototype and evalu ate an array of models. Recent advances allow such algorithms to scale to high d imensions. However, a central question remains: How to specify an expressive var iational distribution that maintains efficient computation? To address this, we develop hierarchical variational models (HVMs). HVMs augment a variational appro ximation with a prior on its parameters, which allows it to capture complex stru cture for both discrete and continuous latent variables. The algorithm we develo p is black box, can be used for any HVM, and has the same computational efficien cy as the original approximation. We study HVMs on a variety of deep discrete la tent variable models. HVMs generalize other expressive variational distributions and maintains higher fidelity to the posterior.
****

Hierarchical Span-Based Conditional Random Fields for Labeling and Segmenting Ev ents in Wearable Sensor Data Streams
Roy Adams, Nazir Saleheen, Edison Thomaz, Abhinav Parate, Santosh Kumar, Benjami n Marlin
The field of mobile health (mHealth) has the potential to yield new insights int o health and behavior through the analysis of continuously recorded data from we arable health and activity sensors. In this paper, we present a hierarchical spa n-based conditional random field model for the key problem of jointly detecting discrete events in such sensor data streams and segmenting these events into hig h-level activity sessions. Our model includes higher-order cardinality factors a nd inter-event duration factors to capture domain-specific structure in the labe l space. We show that our model supports exact MAP inference in quadratic time v ia dynamic programming, which we leverage to perform learning in the structured support vector machine framework. We apply the model to the problems of smoking and eating detection using four real data sets. Our results show statistically s ignificant improvements in segmentation performance relative to a hierarchical p airwise CRF.
****

Binary embeddings with structured hashed projections
Anna Choromanska, Krzysztof Choromanski, Mariusz Bojarski, Tony Jebara, Sanjiv K umar, Yann LeCun
We consider the hashing mechanism for constructing binary embeddings, that invol ves pseudo-random projections followed by nonlinear (sign function) mappings. Th e pseudo-random projection is described by a matrix, where not all entries are i ndependent random variables but instead a fixed “budget of randomness” is distri buted across the matrix. Such matrices can be edfficiently stored in sub-quadrat ic or even linear space, provide reduction in randomness usage (i.e. number of r equired random values), and very often lead to computational speed ups. We prove several theoretical results showing that projections via various structured mat

rices followed by nonlinear mappings accurately preserve the angular distance between input high-dimensional vectors. To the best of our knowledge, these results are the first that give theoretical ground for the use of general structured matrices in the nonlinear setting. We empirically verify our theoretical findings and show the dependence of learning via structured hashed projections on the performance of neural network as well as nearest neighbor classifier.
****

A Variational Analysis of Stochastic Gradient Algorithms
Stephan Mandt, Matthew Hoffman, David Blei
Stochastic Gradient Descent (SGD) is an important algorithm in machine learning. With constant learning rates, it is a stochastic process that, after an initial phase of convergence, generates samples from a stationary distribution. We show that SGD with constant rates can be effectively used as an approximate posterior inference algorithm for probabilistic modeling. Specifically, we show how to adjust the tuning parameters of SGD such as to match the resulting stationary distribution to the posterior. This analysis rests on interpreting SGD as a continuous-time stochastic process and then minimizing the Kullback-Leibler divergence between its stationary distribution and the target posterior. (This is in the spirit of variational inference.) In more detail, we model SGD as a multivariate Ornstein-Uhlenbeck process and then use properties of this process to derive the optimal parameters. This theoretical framework also connects SGD to modern scalable inference algorithms; we analyze the recently proposed stochastic gradient Fisher scoring under this perspective. We demonstrate that SGD with properly chosen constant rates gives a new way to optimize hyperparameters in probabilistic models.
****

Adaptive Sampling for SGD by Exploiting Side Information
Siddharth Gopal
This paper proposes a new mechanism for sampling training instances for stochastic gradient descent (SGD) methods by exploiting any side-information associated with the instances (for e.g. class-labels) to improve convergence. Previous methods have either relied on sampling from a distribution defined over training instances or from a static distribution that fixed before training. This results in two problems a) any distribution that is set apriori is independent of how the optimization progresses and b) maintaining a distribution over individual instances could be infeasible in large-scale scenarios. In this paper, we exploit the side information associated with the instances to tackle both problems. More specifically, we maintain a distribution over classes (instead of individual instances) that is adaptively estimated during the course of optimization to give the maximum reduction in the variance of the gradient. Intuitively, we sample more from those regions in space that have a \textitlarger gradient contribution. Our experiments on highly multiclass datasets show that our proposal converge significantly faster than existing techniques.
****

Learning from Multiway Data: Simple and Efficient Tensor Regression
Rose Yu, Yan Liu
Tensor regression has shown to be advantageous in learning tasks with multi-directional relatedness. Given massive multiway data, traditional methods are often too slow to operate on or suffer from memory bottleneck. In this paper, we introduce subsampled tensor projected gradient to solve the problem. Our algorithm is impressively simple and efficient. It is built upon projected gradient method with fast tensor power iterations, leveraging randomized sketching for further acceleration. Theoretical analysis shows that our algorithm converges to the correct solution in fixed number of iterations. The memory requirement grows linearly with the size of the problem. We demonstrate superior empirical performance on both multi-linear multi-task learning and spatio-temporal applications.
****

A Distributed Variational Inference Framework for Unifying Parallel Sparse Gaussian Process Regression Models
Trong Nghia Hoang, Quang Minh Hoang, Bryan Kian Hsiang Low

This paper presents a novel distributed variational inference framework that unifies many parallel sparse Gaussian process regression (SGPR) models for scalable hyperparameter learning with big data. To achieve this, our framework exploits a structure of correlated noise process model that represents the observation noises as a finite realization of a high-order Gaussian Markov random process. By varying the Markov order and covariance function for the noise process model, different variational SGPR models result. This consequently allows the correlation structure of the noise process model to be characterized for which a particular variational SGPR model is optimal. We empirically evaluate the predictive performance and scalability of the distributed variational SGPR models unified by our framework on two real-world datasets.

****

Online Stochastic Linear Optimization under One-bit Feedback

Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, Zhi-hua Zhou

In this paper, we study a special bandit setting of online stochastic linear optimization, where only one-bit of information is revealed to the learner at each round. This problem has found many applications including online advertisement and online recommendation. We assume the binary feedback is a random variable generated from the logit model, and aim to minimize the regret defined by the unknown linear function. Although the existing method for generalized linear bandit can be applied to our problem, the high computational cost makes it impractical for real-world applications. To address this challenge, we develop an efficient online learning algorithm by exploiting particular structures of the observation model. Specifically, we adopt online Newton step to estimate the unknown parameter and derive a tight confidence region based on the exponential concavity of the logistic loss. Our analysis shows that the proposed algorithm achieves a regret bound of O(d\sqrtT), which matches the optimal result of stochastic linear bandits.

****

Adaptive Algorithms for Online Convex Optimization with Long-term Constraints

Rodolphe Jenatton, Jim Huang, Cedric Archambeau

We present an adaptive online gradient descent algorithm to solve online convex optimization problems with long-term constraints, which are constraints that need to be satisfied when accumulated over a finite number of rounds T, but can be violated in intermediate rounds. For some user-defined trade-off parameter βin (0, 1), the proposed algorithm achieves cumulative regret bounds of O(T^maxβ,1_β) and O(T^1_β/2), respectively for the loss and the constraint violations. Our results hold for convex losses, can handle arbitrary convex constraints and rely on a single computationally efficient algorithm. Our contributions improve over the best known cumulative regret bounds of Mahdavi et al. (2012), which are respectively O(T^1/2) and O(T^3/4) for general convex domains, and respectively O(T^2/3) and O(T^2/3) when the domain is further restricted to be a polyhedral set. We supplement the analysis with experiments validating the performance of our algorithm in practice.

****

Actively Learning Hemimetrics with Applications to Eliciting User Preferences

Adish Singla, Sebastian Tschiatschek, Andreas Krause

Motivated by an application of eliciting users' preferences, we investigate the problem of learning hemimetrics, i.e., pairwise distances among a set of n items that satisfy triangle inequalities and non-negativity constraints. In our application, the (asymmetric) distances quantify private costs a user incurs when substituting one item by another. We aim to learn these distances (costs) by asking the users whether they are willing to switch from one item to another for a given incentive offer. Without exploiting structural constraints of the hemimetric polytope, learning the distances between each pair of items requires Θ(n^2) queries. We propose an active learning algorithm that substantially reduces this sample complexity by exploiting the structural constraints on the version space of hemimetrics. Our proposed algorithm achieves provably-optimal sample complexity for various instances of the task. For example, when the items are embedded into K tight clusters, the sample complexity of our algorithm reduces to O(n K). Ext

ensive experiments on a restaurant recommendation data set support the conclusio
ns of our theoretical analysis.
****

Learning Simple Algorithms from Examples
Wojciech Zaremba, Tomas Mikolov, Armand Joulin, Rob Fergus
We present an approach for learning simple algorithms such as copying, multi-dig
it addition and single digit multiplication directly from examples. Our framewor
k consists of a set of interfaces, accessed by a controller. Typical interfaces
are 1-D tapes or 2-D grids that hold the input and output data. For the controll
er, we explore a range of neural network-based models which vary in their abilit
y to abstract the underlying algorithm from training instances and generalize to
 test examples with many thousands of digits. The controller is trained using Q-
learning with several enhancements and we show that the bottleneck is in the cap
abilities of the controller rather than in the search incurred by Q-learning.
****

Learning Physical Intuition of Block Towers by Example
Adam Lerer, Sam Gross, Rob Fergus
Wooden blocks are a common toy for infants, allowing them to develop motor skill
s and gain intuition about the physical behavior of the world. In this paper, we
 explore the ability of deep feed-forward models to learn such intuitive physics
. Using a 3D game engine, we create small towers of wooden blocks whose stabilit
y is randomized and render them collapsing (or remaining upright). This data all
ows us to train large convolutional network models which can accurately predict
the outcome, as well as estimating the trajectories of the blocks. The models ar
e also able to generalize in two important ways: (i) to new physical scenarios,
e.g. towers with an additional block and (ii) to images of real wooden blocks, w
here it obtains a performance comparable to human subjects.
****

Structure Learning of Partitioned Markov Networks
Song Liu, Taiji Suzuki, Masashi Sugiyama, Kenji Fukumizu
We learn the structure of a Markov Network between two groups of random variable
s from joint observations. Since modelling and learning the full MN structure ma
y be hard, learning the links between two groups directly may be a preferable op
tion. We introduce a novel concept called the \emphpartitioned ratio whose facto
rization directly associates with the Markovian properties of random variables a
cross two groups. A simple one-shot convex optimization procedure is proposed fo
r learning the \emphsparse factorizations of the partitioned ratio and it is the
oretically guaranteed to recover the correct inter-group structure under mild co
nditions. The performance of the proposed method is experimentally compared with
 the state of the art MN structure learning methods using ROC curves. Real appli
cations on analyzing bipartisanship in US congress and pairwise DNA/time-series
alignments are also reported.
****

Tracking Slowly Moving Clairvoyant: Optimal Dynamic Regret of Online Learning wi
th True and Noisy Gradient
Tianbao Yang, Lijun Zhang, Rong Jin, Jinfeng Yi
This work focuses on dynamic regret of online convex optimization that compares
the performance of online learning to a clairvoyant who knows the sequence of lo
ss functions in advance and hence selects the minimizer of the loss function at
each step. By assuming that the clairvoyant moves slowly (i.e., the minimizers c
hange slowly), we present several improved variation-based upper bounds of the d
ynamic regret under the true and noisy gradient feedback, which are \it optimal
in light of the presented lower bounds. The key to our analysis is to explore a
regularity metric that measures the temporal changes in the clairvoyant’s minimi
zers, to which we refer as path variation. Firstly, we present a general lower b
ound in terms of the path variation, and then show that under full information o
r gradient feedback we are able to achieve an optimal dynamic regret. Secondly,
we present a lower bound with noisy gradient feedback and then show that we can
achieve optimal dynamic regrets under a stochastic gradient feedback and two-poi
nt bandit feedback. Moreover, for a sequence of smooth loss functions that admit

a small variation in the gradients, our dynamic regret under the two-point band it feedback matches that is achieved with full information.
****
Beyond CCA: Moment Matching for Multi-View Models
Anastasia Podosinnikova, Francis Bach, Simon Lacoste-Julien
We introduce three novel semi-parametric extensions of probabilistic canonical correlation analysis with identifiability guarantees. We consider moment matching techniques for estimation in these models. For that, by drawing explicit links between the new models and a discrete version of independent component analysis (DICA), we first extend the DICA cumulant tensors to the new discrete version of CCA. By further using a close connection with independent component analysis, we introduce generalized covariance matrices, which can replace the cumulant tensors in the moment matching framework, and, therefore, improve sample complexity and simplify derivations and algorithms significantly. As the tensor power method or orthogonal joint diagonalization are not applicable in the new setting, we use non-orthogonal joint diagonalization techniques for matching the cumulants. We demonstrate performance of the proposed models and estimation techniques on experiments with both synthetic and real datasets.
****
Fast methods for estimating the Numerical rank of large matrices
Shashanka Ubaru, Yousef Saad
We present two computationally inexpensive techniques for estimating the numerical rank of a matrix, combining powerful tools from computational linear algebra. These techniques exploit three key ingredients. The first is to approximate the projector on the non-null invariant subspace of the matrix by using a polynomial filter. Two types of filters are discussed, one based on Hermite interpolation and the other based on Chebyshev expansions. The second ingredient employs stochastic trace estimators to compute the rank of this wanted eigen-projector, which yields the desired rank of the matrix. In order to obtain a good filter, it is necessary to detect a gap between the eigenvalues that correspond to noise and the relevant eigenvalues that correspond to the non-null invariant subspace. The third ingredient of the proposed approaches exploits the idea of spectral density, popular in physics, and the Lanczos spectroscopic method to locate this gap.
****
Unsupervised Deep Embedding for Clustering Analysis
Junyuan Xie, Ross Girshick, Ali Farhadi
Clustering is central to many data-driven application domains and has been studied extensively in terms of distance functions and grouping algorithms. Relatively little work has focused on learning representations for clustering. In this paper, we propose Deep Embedded Clustering (DEC), a method that simultaneously learns feature representations and cluster assignments using deep neural networks. DEC learns a mapping from the data space to a lower-dimensional feature space in which it iteratively optimizes a clustering objective. Our experimental evaluations on image and text corpora show significant improvement over state-of-the-art methods.
****
Efficient Private Empirical Risk Minimization for High-dimensional Learning
Shiva Prasad Kasiviswanathan, Hongxia Jin
Dimensionality reduction is a popular approach for dealing with high dimensional data that leads to substantial computational savings. Random projections are a simple and effective method for universal dimensionality reduction with rigorous theoretical guarantees. In this paper, we theoretically study the problem of differentially private empirical risk minimization in the projected subspace (compressed domain). We ask: is it possible to design differentially private algorithms with small excess risk given access to only projected data? In this paper, we answer this question in affirmative, by showing that for the class of generalized linear functions, given only the projected data and the projection matrix, we can obtain excess risk bounds of $O(w(Theta)^2/3/n^1/3) under eps-differential privacy, and O((w(Theta)/n)^1/2)$ under (eps,delta)-differential privacy, where n is the sample size and w(Theta) is the Gaussian width of the parameter space t

hat we optimize over. A simple consequence of these results is that, for a large class of ERM problems, in the traditional setting (i.e., with access to the original data), under eps-differential privacy, we improve the worst-case risk bounds of Bassily et al. (FOCS 2014).
****
Parameter Estimation for Generalized Thurstone Choice Models
Milan Vojnovic, Seyoung Yun
We consider the maximum likelihood parameter estimation problem for a generalized Thurstone choice model, where choices are from comparison sets of two or more items. We provide tight characterizations of the mean square error, as well as necessary and sufficient conditions for correct classification when each item belongs to one of two classes. These results provide insights into how the estimation accuracy depends on the choice of a generalized Thurstone choice model and the structure of comparison sets. We find that for a priori unbiased structures of comparisons, e.g., when comparison sets are drawn independently and uniformly at random, the number of observations needed to achieve a prescribed estimation accuracy depends on the choice of a generalized Thurstone choice model. For a broad set of generalized Thurstone choice models, which includes all popular instances used in practice, the estimation error is shown to be largely insensitive to the cardinality of comparison sets. On the other hand, we found that there exist generalized Thurstone choice models for which the estimation error decreases much faster with the cardinality of comparison sets.
****
Large-Margin Softmax Loss for Convolutional Neural Networks
Weiyang Liu, Yandong Wen, Zhiding Yu, Meng Yang
Cross-entropy loss together with softmax is arguably one of the most common used supervision components in convolutional neural networks (CNNs). Despite its simplicity, popularity and excellent performance, the component does not explicitly encourage discriminative learning of features. In this paper, we propose a generalized large-margin softmax (L-Softmax) loss which explicitly encourages intra-class compactness and inter-class separability between learned features. Moreover, L-Softmax not only can adjust the desired margin but also can avoid overfitting. We also show that the L-Softmax loss can be optimized by typical stochastic gradient descent. Extensive experiments on four benchmark datasets demonstrate that the deeply-learned features with L-softmax loss become more discriminative, hence significantly boosting the performance on a variety of visual classification and verification tasks.
****
A Random Matrix Approach to Echo-State Neural Networks
Romain Couillet, Gilles Wainrib, Hafiz Tiomoko Ali, Harry Sevi
Recurrent neural networks, especially in their linear version, have provided many qualitative insights on their performance under different configurations. This article provides, through a novel random matrix framework, the quantitative counterpart of these performance results, specifically in the case of echo-state networks. Beyond mere insights, our approach conveys a deeper understanding on the core mechanism under play for both training and testing.
****
Supervised and Semi-Supervised Text Categorization using LSTM for Region Embeddings
Rie Johnson, Tong Zhang
One-hot CNN (convolutional neural network) has been shown to be effective for text categorization (Johnson & Zhang, 2015). We view it as a special case of a general framework which jointly trains a linear model with a non-linear feature generator consisting of ‘text region embedding + pooling’. Under this framework, we explore a more sophisticated region embedding method using Long Short-Term Memory (LSTM). LSTM can embed text regions of variable (and possibly large) sizes, whereas the region size needs to be fixed in a CNN. We seek effective and efficient use of LSTM for this purpose in the supervised and semi-supervised settings. The best results were obtained by combining region embeddings in the form of LSTM and convolution layers trained on unlabeled data. The results indicate that on

this task, embeddings of text regions, which can convey complex concepts, are more useful than embeddings of single words in isolation. We report performances exceeding the previous best results on four benchmark datasets.

****

Optimality of Belief Propagation for Crowdsourced Classification

Jungseul Ok, Sewoong Oh, Jinwoo Shin, Yung Yi

Crowdsourcing systems are popular for solving large-scale labelling tasks with low-paid (or even non-paid) workers. We study the problem of recovering the true labels from noisy crowdsourced labels under the popular Dawid-Skene model. To address this inference problem, several algorithms have recently been proposed, but the best known guarantee is still significantly larger than the fundamental limit. We close this gap under a simple but canonical scenario where each worker is assigned at most two tasks. In particular, we introduce a tighter lower bound on the fundamental limit and prove that Belief Propagation (BP) exactly matches this lower bound. The guaranteed optimality of BP is the strongest in the sense that it is information-theoretically impossible for any other algorithm to correctly la- bel a larger fraction of the tasks. In the general setting, when more than two tasks are assigned to each worker, we establish the dominance result on BP that it outperforms other existing algorithms with known provable guarantees. Experimental results suggest that BP is close to optimal for all regimes considered, while existing state-of-the-art algorithms exhibit suboptimal performances.

****

Stability of Controllers for Gaussian Process Forward Models

Julia Vinogradska, Bastian Bischoff, Duy Nguyen-Tuong, Anne Romer, Henner Schmidt, Jan Peters

Learning control has become an appealing alternative to the derivation of control laws based on classic control theory. However, a major shortcoming of learning control is the lack of performance guarantees which prevents its application in many real-world scenarios. As a step in this direction, we provide a stability analysis tool for controllers acting on dynamics represented by Gaussian processes (GPs). We consider arbitrary Markovian control policies and system dynamics given as (i) the mean of a GP, and (ii) the full GP distribution. For the first case, our tool finds a state space region, where the closed-loop system is provably stable. In the second case, it is well known that infinite horizon stability guarantees cannot exist. Instead, our tool analyzes finite time stability. Empirical evaluations on simulated benchmark problems support our theoretical results.

****

Learning privately from multiparty data

Jihun Hamm, Yingjun Cao, Mikhail Belkin

Learning a classifier from private data distributed across multiple parties is an important problem that has many potential applications. How can we build an accurate and differentially private global classifier by combining locally-trained classifiers from different parties, without access to any party’s private data? We propose to transfer the “knowledge” of the local classifier ensemble by first creating labeled data from auxiliary unlabeled data, and then train a global differentially private classifier. We show that majority voting is too sensitive and therefore propose a new risk weighted by class probabilities estimated from the ensemble. Relative to a non-private solution, our private solution has a generalization error bounded by O(ε^-2 M^-2). This allows strong privacy without performance loss when the number of participating parties M is large, such as in crowdsensing applications. We demonstrate the performance of our framework with realistic tasks of activity recognition, network intrusion detection, and malicious URL detection.

****

Network Morphism

Tao Wei, Changhu Wang, Yong Rui, Chang Wen Chen

We present a systematic study on how to morph a well-trained neural network to a new one so that its network function can be completely preserved. We define thi

s as network morphism in this research. After morphing a parent network, the chi ld network is expected to inherit the knowledge from its parent network and also has the potential to continue growing into a more powerful one with much shorte ned training time. The first requirement for this network morphism is its abilit y to handle diverse morphing types of networks, including changes of depth, widt h, kernel size, and even subnet. To meet this requirement, we first introduce th e network morphism equations, and then develop novel morphing algorithms for all these morphing types for both classic and convolutional neural networks. The se cond requirement is its ability to deal with non-linearity in a network. We prop ose a family of parametric-activation functions to facilitate the morphing of an y continuous non-linear activation neurons. Experimental results on benchmark da tasets and typical neural networks demonstrate the effectiveness of the proposed network morphism scheme.

****

A Kronecker-factored approximate Fisher matrix for convolution layers

Roger Grosse, James Martens

Second-order optimization methods such as natural gradient descent have the pote ntial to speed up training of neural networks by correcting for the curvature of the loss function. Unfortunately, the exact natural gradient is impractical to compute for large models, and most approximations either require an expensive it erative procedure or make crude approximations to the curvature. We present Kron ecker Factors for Convolution (KFC), a tractable approximation to the Fisher mat rix for convolutional networks based on a structured probabilistic model for the distribution over backpropagated derivatives. Similarly to the recently propose d Kronecker-Factored Approximate Curvature (K-FAC), each block of the approximat e Fisher matrix decomposes as the Kronecker product of small matrices, allowing for efficient inversion. KFC captures important curvature information while stil l yielding comparably efficient updates to stochastic gradient descent (SGD). We show that the updates are invariant to commonly used reparameterizations, such as centering of the activations. In our experiments, approximate natural gradien t descent with KFC was able to train convolutional networks several times faster than carefully tuned SGD. Furthermore, it was able to train the networks in 10-20 times fewer iterations than SGD, suggesting its potential applicability in a distributed setting.

****

Experimental Design on a Budget for Sparse Linear Models and Applications

Sathya Narayanan Ravi, Vamsi Ithapu, Sterling Johnson, Vikas Singh

Budget constrained optimal design of experiments is a classical problem in stati stics. Although the optimal design literature is very mature, few efficient stra tegies are available when these design problems appear in the context of sparse linear models commonly encountered in high dimensional machine learning and stat istics. In this work, we study experimental design for the setting where the und erlying regression model is characterized by a \ell_1-regularized linear functio n. We propose two novel strategies: the first is motivated geometrically whereas the second is algebraic in nature. We obtain tractable algorithms for this prob lem and also hold for a more general class of sparse linear models. We perform a n extensive set of experiments, on benchmarks and a large multi-site neuroscienc e study, showing that the proposed models are effective in practice. The latter experiment suggests that these ideas may play a small role in informing enrollme nt strategies for similar scientific studies in the short-to-medium term future.

****

Minding the Gaps for Block Frank-Wolfe Optimization of Structured SVMs

Anton Osokin, Jean-Baptiste Alayrac, Isabella Lukasewitz, Puneet Dokania, Simon Lacoste-Julien

In this paper, we propose several improvements on the block-coordinate Frank-Wol fe (BCFW) algorithm from Lacoste-Julien et al. (2013) recently used to optimize the structured support vector machine (SSVM) objective in the context of structu red prediction, though it has wider applications. The key intuition behind our i mprovements is that the estimates of block gaps maintained by BCFW reveal the bl ock suboptimality that can be used as an *adaptive* criterion. First, we sample

objects at each iteration of BCFW in an adaptive non-uniform way via gap-based s
ampling. Second, we incorporate pairwise and away-step variants of Frank-Wolfe i
nto the block-coordinate setting. Third, we cache oracle calls with a cache-hit
criterion based on the block gaps. Fourth, we provide the first method to comput
e an approximate regularization path for SSVM. Finally, we provide an exhaustive
 empirical evaluation of all our methods on four structured prediction datasets.
****

Exact Exponent in Optimal Rates for Crowdsourcing
Chao Gao, Yu Lu, Dengyong Zhou
Crowdsourcing has become a popular tool for labeling large datasets. This paper
studies the optimal error rate for aggregating crowdsourced labels provided by a
 collection of amateur workers. Under the Dawid-Skene probabilistic model, we es
tablish matching upper and lower bounds with an exact exponent mI(\pi), where m
is the number of workers and I(\pi) is the average Chernoff information that cha
racterizes the workers' collective ability. Such an exact characterization of th
e error exponent allows us to state a precise sample size requirement m \ge \fra
c1I(\pi)\log\frac1ε in order to achieve an εmisclassification error. In addition
, our results imply optimality of various forms of EM algorithms given accurate
initializers of the model parameters.
****

Augmenting Supervised Neural Networks with Unsupervised Objectives for Large-sca
le Image Classification
Yuting Zhang, Kibok Lee, Honglak Lee
Unsupervised learning and supervised learning are key research topics in deep le
arning. However, as high-capacity supervised neural networks trained with a larg
e amount of labels have achieved remarkable success in many computer vision task
s, the availability of large-scale labeled images reduced the significance of un
supervised learning. Inspired by the recent trend toward revisiting the importan
ce of unsupervised learning, we investigate joint supervised and unsupervised le
arning in a large-scale setting by augmenting existing neural networks with deco
ding pathways for reconstruction. First, we demonstrate that the intermediate ac
tivations of pretrained large-scale classification networks preserve almost all
the information of input images except a portion of local spatial details. Then,
 by end-to-end training of the entire augmented architecture with the reconstruc
tive objective, we show improvement of the network performance for supervised ta
sks. We evaluate several variants of autoencoders, including the recently propos
ed "what-where" autoencoder that uses the encoder pooling switches, to study the
 importance of the architecture design. Taking the 16-layer VGGNet trained under
 the ImageNet ILSVRC 2012 protocol as a strong baseline for image classification
, our methods improve the validation-set accuracy by a noticeable margin.
****

Online Low-Rank Subspace Clustering by Basis Dictionary Pursuit
Jie Shen, Ping Li, Huan Xu
Low-Rank Representation (LRR) has been a significant method for segmenting data
that are generated from a union of subspaces. It is also known that solving LRR
is challenging in terms of time complexity and memory footprint, in that the siz
e of the nuclear norm regularized matrix is n-by-n (where n is the number of sam
ples). In this paper, we thereby develop a novel online implementation of LRR th
at reduces the memory cost from O(n^2) to O(pd), with p being the ambient dimens
ion and d being some estimated rank (d < p < n). We also establish the theoretic
al guarantee that the sequence of solutions produced by our algorithm converges
to a stationary point of the expected loss function asymptotically. Extensive ex
periments on synthetic and realistic datasets further substantiate that our algo
rithm is fast, robust and memory efficient.
****

A Self-Correcting Variable-Metric Algorithm for Stochastic Optimization
Frank Curtis
An algorithm for stochastic (convex or nonconvex) optimization is presented. The
 algorithm is variable-metric in the sense that, in each iteration, the step is
computed through the product of a symmetric positive definite scaling matrix and

a stochastic (mini-batch) gradient of the objective function, where the sequence of scaling matrices is updated dynamically by the algorithm. A key feature of the algorithm is that it does not overly restrict the manner in which the scaling matrices are updated. Rather, the algorithm exploits fundamental self-correcting properties of BFGS-type updating—properties that have been over-looked in other attempts to devise quasi-Newton methods for stochastic optimization. Numerical experiments illustrate that the method and a limited memory variant of it are stable and outperform (mini-batch) stochastic gradient and other quasi-Newton methods when employed to solve a few machine learning problems.

****

Stochastic Quasi-Newton Langevin Monte Carlo

Umut Simsekli, Roland Badeau, Taylan Cemgil, Gaël Richard

Recently, Stochastic Gradient Markov Chain Monte Carlo (SG-MCMC) methods have been proposed for scaling up Monte Carlo computations to large data problems. Whilst these approaches have proven useful in many applications, vanilla SG-MCMC might suffer from poor mixing rates when random variables exhibit strong couplings under the target densities or big scale differences. In this study, we propose a novel SG-MCMC method that takes the local geometry into account by using ideas from Quasi-Newton optimization methods. These second order methods directly approximate the inverse Hessian by using a limited history of samples and their gradients. Our method uses dense approximations of the inverse Hessian while keeping the time and memory complexities linear with the dimension of the problem. We provide a formal theoretical analysis where we show that the proposed method is asymptotically unbiased and consistent with the posterior expectations. We illustrate the effectiveness of the approach on both synthetic and real datasets. Our experiments on two challenging applications show that our method achieves fast convergence rates similar to Riemannian approaches while at the same time having low computational requirements similar to diagonal preconditioning approaches.

****

Doubly Robust Off-policy Value Evaluation for Reinforcement Learning

Nan Jiang, Lihong Li

We study the problem of off-policy value evaluation in reinforcement learning (RL), where one aims to estimate the value of a new policy based on data collected by a different policy. This problem is often a critical step when applying RL to real-world problems. Despite its importance, existing general methods either have uncontrolled bias or suffer high variance. In this work, we extend the doubly robust estimator for bandits to sequential decision-making problems, which gets the best of both worlds: it is guaranteed to be unbiased and can have a much lower variance than the popular importance sampling estimators. We demonstrate the estimator’s accuracy in several benchmark problems, and illustrate its use as a subroutine in safe policy improvement. We also provide theoretical results on the inherent hardness of the problem, and show that our estimator can match the lower bound in certain scenarios.

****

Fast Rate Analysis of Some Stochastic Optimization Algorithms

Chao Qu, Huan Xu, Chong Ong

In this paper, we revisit three fundamental and popular stochastic optimization algorithms (namely, Online Proximal Gradient, Regularized Dual Averaging method and ADMM with online proximal gradient) and analyze their convergence speed under conditions weaker than those in literature. In particular, previous works showed that these algorithms converge at a rate of O (\ln T/T) when the loss function is strongly convex, and O (1 /\sqrtT) in the weakly convex case. In contrast, we relax the strong convexity assumption of the loss function, and show that the algorithms converge at a rate O (\ln T/T) if the \em expectation of the loss function is \em locally strongly convex. This is a much weaker assumption and is satisfied by many practical formulations including Lasso and Logistic Regression. Our analysis thus extends the applicability of these three methods, as well as provides a general recipe for improving analysis of convergence rate for stochastic and online optimization algorithms.

****

Fast k-Nearest Neighbour Search via Dynamic Continuous Indexing
Ke Li, Jitendra Malik
Existing methods for retrieving k-nearest neighbours suffer from the curse of dimensionality. We argue this is caused in part by inherent deficiencies of space partitioning, which is the underlying strategy used by most existing methods. We devise a new strategy that avoids partitioning the vector space and present a novel randomized algorithm that runs in time linear in dimensionality of the space and sub-linear in the intrinsic dimensionality and the size of the dataset and takes space constant in dimensionality of the space and linear in the size of the dataset. The proposed algorithm allows fine-grained control over accuracy and speed on a per-query basis, automatically adapts to variations in data density, supports dynamic updates to the dataset and is easy-to-implement. We show appealing theoretical properties and demonstrate empirically that the proposed algorithm outperforms locality-sensitivity hashing (LSH) in terms of approximation quality, speed and space efficiency.
****
Smooth Imitation Learning for Online Sequence Prediction
Hoang Le, Andrew Kang, Yisong Yue, Peter Carr
We study the problem of smooth imitation learning for online sequence prediction, where the goal is to train a policy that can smoothly imitate demonstrated behavior in a dynamic and continuous environment in response to online, sequential context input. Since the mapping from context to behavior is often complex, we take a learning reduction approach to reduce smooth imitation learning to a regression problem using complex function classes that are regularized to ensure smoothness. We present a learning meta-algorithm that achieves fast and stable convergence to a good policy. Our approach enjoys several attractive properties, including being fully deterministic, employing an adaptive learning rate that can provably yield larger policy improvements compared to previous approaches, and the ability to ensure stable convergence. Our empirical results demonstrate significant performance gains over previous approaches.
****
Community Recovery in Graphs with Locality
Yuxin Chen, Govinda Kamath, Changho Suh, David Tse
Motivated by applications in domains such as social networks and computational biology, we study the problem of community recovery in graphs with locality. In this problem, pairwise noisy measurements of whether two nodes are in the same community or different communities come mainly or exclusively from nearby nodes rather than uniformly sampled between all node pairs, as in most existing models. We present two algorithms that run nearly linearly in the number of measurements and which achieve the information limits for exact recovery.
****
Variance Reduction for Faster Non-Convex Optimization
Zeyuan Allen-Zhu, Elad Hazan
We consider the fundamental problem in non-convex optimization of efficiently reaching a stationary point. In contrast to the convex case, in the long history of this basic problem, the only known theoretical results on first-order non-convex optimization remain to be full gradient descent that converges in O(1/\varepsilon) iterations for smooth objectives, and stochastic gradient descent that converges in O(1/\varepsilon^2) iterations for objectives that are sum of smooth functions. We provide the first improvement in this line of research. Our result is based on the variance reduction trick recently introduced to convex optimization, as well as a brand new analysis of variance reduction that is suitable for non-convex optimization. For objectives that are sum of smooth functions, our first-order minibatch stochastic method converges with an O(1/\varepsilon) rate, and is faster than full gradient descent by Ω(n^1/3). We demonstrate the effectiveness of our methods on empirical risk minimizations with non-convex loss functions and training neural nets.
****
Loss factorization, weakly supervised learning and label noise robustness
Giorgio Patrini, Frank Nielsen, Richard Nock, Marcello Carioni

We prove that the empirical risk of most well-known loss functions factors into a linear term aggregating all labels with a term that is label free, and can further be expressed by sums of the same loss. This holds true even for non-smooth, non-convex losses and in any RKHS. The first term is a (kernel) mean operator — the focal quantity of this work — which we characterize as the sufficient statistic for the labels. The result tightens known generalization bounds and sheds new light on their interpretation. Factorization has a direct application on weakly supervised learning. In particular, we demonstrate that algorithms like SGD and proximal methods can be adapted with minimal effort to handle weak supervision, once the mean operator has been estimated. We apply this idea to learning with asymmetric noisy labels, connecting and extending prior work. Furthermore, we show that most losses enjoy a data-dependent (by the mean operator) form of noise robustness, in contrast with known negative results.
****

Analysis of Deep Neural Networks with Extended Data Jacobian Matrix
Shengjie Wang, Abdel-rahman Mohamed, Rich Caruana, Jeff Bilmes, Matthai Plilipose, Matthew Richardson, Krzysztof Geras, Gregor Urban, Ozlem Aslan
Deep neural networks have achieved great successes on various machine learning tasks, however, there are many open fundamental questions to be answered. In this paper, we tackle the problem of quantifying the quality of learned wights of different networks with possibly different architectures, going beyond considering the final classification error as the only metric. We introduce \emphExtended Data Jacobian Matrix to help analyze properties of networks of various structures, finding that, the spectrum of the extended data jacobian matrix is a strong discriminating factor for networks of different structures and performance. Based on such observation, we propose a novel regularization method, which manages to improve the network performance comparably to dropout, which in turn verifies the observation.
****

Doubly Decomposing Nonparametric Tensor Regression
Masaaki Imaizumi, Kohei Hayashi
Nonparametric extension of tensor regression is proposed. Nonlinearity in a high-dimensional tensor space is broken into simple local functions by incorporating low-rank tensor decomposition. Compared to naive nonparametric approaches, our formulation considerably improves the convergence rate of estimation while maintaining consistency with the same function class under specific conditions. To estimate local functions, we develop a Bayesian estimator with the Gaussian process prior. Experimental results show its theoretical properties and high performance in terms of predicting a summary statistic of a real complex network.
****

Hyperparameter optimization with approximate gradient
Fabian Pedregosa
Most models in machine learning contain at least one hyperparameter to control for model complexity. Choosing an appropriate set of hyperparameters is both crucial in terms of model accuracy and computationally challenging. In this work we propose an algorithm for the optimization of continuous hyperparameters using inexact gradient information. An advantage of this method is that hyperparameters can be updated before model parameters have fully converged. We also give sufficient conditions for the global convergence of this method, based on regularity conditions of the involved functions and summability of errors. Finally, we validate the empirical performance of this method on the estimation of regularization constants of L2-regularized logistic regression and kernel Ridge regression. Empirical benchmarks indicate that our approach is highly competitive with respect to state of the art methods.
****

SDCA without Duality, Regularization, and Individual Convexity
Shai Shalev-Shwartz
Stochastic Dual Coordinate Ascent is a popular method for solving regularized loss minimization for the case of convex losses. We describe variants of SDCA that do not require explicit regularization and do not rely on duality. We prove lin

ear convergence rates even if individual loss functions are non-convex, as long as the expected loss is strongly convex.

****

Heteroscedastic Sequences: Beyond Gaussianity

Oren Anava, Shie Mannor

We address the problem of sequential prediction in the heteroscedastic setting, when both the signal and its variance are assumed to depend on explanatory variables. By applying regret minimization techniques, we devise an efficient online learning algorithm for the problem, without assuming that the error terms comply with a specific distribution. We show that our algorithm can be adjusted to provide confidence bounds for its predictions, and provide an application to ARCH models. The theoretic results are corroborated by an empirical study.

****

A Neural Autoregressive Approach to Collaborative Filtering

Yin Zheng, Bangsheng Tang, Wenkui Ding, Hanning Zhou

This paper proposes CF-NADE, a neural autoregressive architecture for collaborative filtering (CF) tasks, which is inspired by the Restricted Boltzmann Machine (RBM) based CF model and the Neural Autoregressive Distribution Estimator (NADE). We first describe the basic CF-NADE model for CF tasks. Then we propose to improve the model by sharing parameters between different ratings. A factored version of CF-NADE is also proposed for better scalability. Furthermore, we take the ordinal nature of the preferences into consideration and propose an ordinal cost to optimize CF-NADE, which shows superior performance. Finally, CF-NADE can be extended to a deep model, with only moderately increased computational complexity. Experimental results show that CF-NADE with a single hidden layer beats all previous state-of-the-art methods on MovieLens 1M, MovieLens 10M, and Netflix datasets, and adding more hidden layers can further improve the performance.

****

On the Quality of the Initial Basin in Overspecified Neural Networks

Itay Safran, Ohad Shamir

Deep learning, in the form of artificial neural networks, has achieved remarkable practical success in recent years, for a variety of difficult machine learning applications. However, a theoretical explanation for this remains a major open problem, since training neural networks involves optimizing a highly non-convex objective function, and is known to be computationally hard in the worst case. In this work, we study the \emphgeometric structure of the associated non-convex objective function, in the context of ReLU networks and starting from a random initialization of the network parameters. We identify some conditions under which it becomes more favorable to optimization, in the sense of (i) High probability of initializing at a point from which there is a monotonically decreasing path to a global minimum; and (ii) High probability of initializing at a basin (suitably defined) with a small minimal objective value. A common theme in our results is that such properties are more likely to hold for larger (“overspecified”) networks, which accords with some recent empirical and theoretical observations.

****

Primal-Dual Rates and Certificates

Celestine Dünner, Simone Forte, Martin Takac, Martin Jaggi

We propose an algorithm-independent framework to equip existing optimization methods with primal-dual certificates. Such certificates and corresponding rate of convergence guarantees are important for practitioners to diagnose progress, in particular in machine learning applications. We obtain new primal-dual convergence rates, e.g., for the Lasso as well as many L1, Elastic Net, group Lasso and TV-regularized problems. The theory applies to any norm-regularized generalized linear model. Our approach provides efficiently computable duality gaps which are globally defined, without modifying the original problems in the region of interest.

****

Minimizing the Maximal Loss: How and Why

Shai Shalev-Shwartz, Yonatan Wexler

A commonly used learning rule is to approximately minimize the \emphaverage loss

over the training set. Other learning algorithms, such as AdaBoost and hard-SVM , aim at minimizing the \emphmaximal loss over the training set. The average los s is more popular, particularly in deep learning, due to three main reasons. Fir st, it can be conveniently minimized using online algorithms, that process few e xamples at each iteration. Second, it is often argued that there is no sense to minimize the loss on the training set too much, as it will not be reflected in t he generalization loss. Last, the maximal loss is not robust to outliers. In thi s paper we describe and analyze an algorithm that can convert any online algorit hm to a minimizer of the maximal loss. We show, theoretically and empirically, t hat in some situations better accuracy on the training set is crucial to obtain good performance on unseen examples. Last, we propose robust versions of the app roach that can handle outliers.

****

The Information-Theoretic Requirements of Subspace Clustering with Missing Data

Daniel Pimentel-Alarcon, Robert Nowak

Subspace clustering with missing data (SCMD) is a useful tool for analyzing inco mplete datasets. Let d be the ambient dimension, and r the dimension of the subs paces. Existing theory shows that Nk = O(r d) columns per subspace are necessary for SCMD, and Nk =O(min d^(log d), d^(r+1)) are sufficient. We close this gap, showing that Nk =O(r d) is also sufficient. To do this we derive deterministic sampling conditions for SCMD, which give precise information theoretic requireme nts and determine sampling regimes. These results explain the performance of SCM D algorithms from the literature. Finally, we give a practical algorithm to cert ify the output of any SCMD method deterministically.

****

Online Learning with Feedback Graphs Without the Graphs

Alon Cohen, Tamir Hazan, Tomer Koren

We study an online learning framework introduced by Mannor and Shamir (2011) in which the feedback is specified by a graph, in a setting where the graph may var y from round to round and is \emphnever fully revealed to the learner. We show a large gap between the adversarial and the stochastic cases. In the adversarial case, we prove that even for dense feedback graphs, the learner cannot improve u pon a trivial regret bound obtained by ignoring any additional feedback besides her own loss. In contrast, in the stochastic case we give an algorithm that achi eves \widetildeΘ(\sqrtαT) regret over T rounds, provided that the independence n umbers of the hidden feedback graphs are at most α. We also extend our results t o a more general feedback model, in which the learner does not necessarily obser ve her own loss, and show that, even in simple cases, concealing the feedback gr aphs might render the problem unlearnable.

****

PAC learning of Probabilistic Automaton based on the Method of Moments

Hadrien Glaude, Olivier Pietquin

Probabilitic Finite Automata (PFA) are generative graphical models that define d istributions with latent variables over finite sequences of symbols, a.k.a. stoc hastic languages. Traditionally, unsupervised learning of PFA is performed throu gh algorithms that iteratively improves the likelihood like the Expectation-Maxi mization (EM) algorithm. Recently, learning algorithms based on the so-called Me thod of Moments (MoM) have been proposed as a much faster alternative that comes with PAC-style guarantees. However, these algorithms do not ensure the learnt a utomata to model a proper distribution, limiting their applicability and prevent ing them to serve as an initialization to iterative algorithms. In this paper, w e propose a new MoM-based algorithm with PAC-style guarantees that learns automa ta defining proper distributions. We assess its performances on synthetic proble ms from the PAutomaC challenge and real datasets extracted from Wikipedia agains t previous MoM-based algorithms and EM algorithm.

****

Estimating Structured Vector Autoregressive Models

Igor Melnyk, Arindam Banerjee

While considerable advances have been made in estimating high-dimensional struct ured models from independent data using Lasso-type models, limited progress has

been made for settings when the samples are dependent. We consider estimating st ructured VAR (vector auto-regressive model), where the structure can be captured by any suitable norm, e.g., Lasso, group Lasso, order weighted Lasso, etc. In V AR setting with correlated noise, although there is strong dependence over time and covariates, we establish bounds on the non-asymptotic estimation error of st ructured VAR parameters. The estimation error is of the same order as that of th e corresponding Lasso-type estimator with independent samples, and the analysis holds for any norm. Our analysis relies on results in generic chaining, sub-expo nential martingales, and spectral representation of VAR models. Experimental res ults on synthetic and real data with a variety of structures are presented, vali dating theoretical results.
****

Mixing Rates for the Alternating Gibbs Sampler over Restricted Boltzmann Machine s and Friends
Christopher Tosh
Alternating Gibbs sampling is a modification of classical Gibbs sampling where s everal variables are simultaneously sampled from their joint conditional distrib ution. In this work, we investigate the mixing rate of alternating Gibbs samplin g with a particular emphasis on Restricted Boltzmann Machines (RBMs) and variant s.
****

Polynomial Networks and Factorization Machines: New Insights and Efficient Train ing Algorithms
Mathieu Blondel, Masakazu Ishihata, Akinori Fujino, Naonori Ueda
Polynomial networks and factorization machines are two recently-proposed models that can efficiently use feature interactions in classification and regression t asks. In this paper, we revisit both models from a unified perspective. Based on this new view, we study the properties of both models and propose new efficient training algorithms. Key to our approach is to cast parameter learning as a low -rank symmetric tensor estimation problem, which we solve by multi-convex optimi zation. We demonstrate our approach on regression and recommender system tasks.
****

A New PAC-Bayesian Perspective on Domain Adaptation
Pascal Germain, Amaury Habrard, François Laviolette, Emilie Morvant
We study the issue of PAC-Bayesian domain adaptation: We want to learn, from a s ource domain, a majority vote model dedicated to a target one. Our theoretical c ontribution brings a new perspective by deriving an upper-bound on the target ri sk where the distributions’ divergence - expressed as a ratio - controls the tra de-off between a source error measure and the target voters’ disagreement. Our b ound suggests that one has to focus on regions where the source data is informat ive. From this result, we derive a PAC-Bayesian generalization bound, and specia lize it to linear classifiers. Then, we infer a learning algorithm and perform e xperiments on real data.
****

Correlation Clustering and Biclustering with Locally Bounded Errors
Gregory Puleo, Olgica Milenkovic
We consider a generalized version of the correlation clustering problem, defined as follows. Given a complete graph G whose edges are labeled with + or -, we wi sh to partition the graph into clusters while trying to avoid errors: + edges be tween clusters or - edges within clusters. Classically, one seeks to minimize th e total number of such errors. We introduce a new framework that allows the obje ctive to be a more general function of the number of errors at each vertex (for example, we may wish to minimize the number of errors at the worst vertex) and p rovide a rounding algorithm which converts “fractional clusterings” into discret e clusterings while causing only a constant-factor blowup in the number of error s at each vertex. This rounding algorithm yields constant-factor approximation a lgorithms for the discrete problem under a wide variety of objective functions.
****

PAC Lower Bounds and Efficient Algorithms for The Max K-Armed Bandit Problem
Yahel David, Nahum Shimkin

We consider the Max K-Armed Bandit problem, where a learning agent is faced with several stochastic arms, each a source of i.i.d. rewards of unknown distribution. At each time step the agent chooses an arm, and observes the reward of the obtained sample. Each sample is considered here as a separate item with the reward designating its value, and the goal is to find an item with the highest possible value. Our basic assumption is a known lower bound on the \em tail function of the reward distributions. Under the PAC framework, we provide a lower bound on the sample complexity of any (ε,δ)-correct algorithm, and propose an algorithm that attains this bound up to logarithmic factors. We provide an analysis of the robustness of the proposed algorithm to the model assumptions, and further compare its performance to the simple non-adaptive variant, in which the arms are chosen randomly at each stage.
****

A Comparative Analysis and Study of Multiview CNN Models for Joint Object Categorization and Pose Estimation
Mohamed Elhoseiny, Tarek El-Gaaly, Amr Bakry, Ahmed Elgammal
In the Object Recognition task, there exists a dichotomy between the categorization of objects and estimating object pose, where the former necessitates a view-invariant representation, while the latter requires a representation capable of capturing pose information over different categories of objects. With the rise of deep architectures, the prime focus has been on object category recognition. Deep learning methods have achieved wide success in this task. In contrast, object pose estimation using these approaches has received relatively less attention. In this work, we study how Convolutional Neural Networks (CNN) architectures can be adapted to the task of simultaneous object recognition and pose estimation. We investigate and analyze the layers of various CNN models and extensively compare between them with the goal of discovering how the layers of distributed representations within CNNs represent object pose information and how this contradicts with object category representations. We extensively experiment on two recent large and challenging multi-view datasets and we achieve better than the state-of-the-art.
****

BASC: Applying Bayesian Optimization to the Search for Global Minima on Potential Energy Surfaces
Shane Carr, Roman Garnett, Cynthia Lo
We present a novel application of Bayesian optimization to the field of surface science: rapidly and accurately searching for the global minimum on potential energy surfaces. Controlling molecule-surface interactions is key for applications ranging from environmental catalysis to gas sensing. We present pragmatic techniques, including exploration/exploitation scheduling and a custom covariance kernel that encodes the properties of our objective function. Our method, the Bayesian Active Site Calculator (BASC), outperforms differential evolution and constrained minima hopping – two state-of-the-art approaches – in trial examples of carbon monoxide adsorption on a hematite substrate, both with and without a defect.
****

On the Iteration Complexity of Oblivious First-Order Optimization Algorithms
Yossi Arjevani, Ohad Shamir
We consider a broad class of first-order optimization algorithms which are \emph oblivious, in the sense that their step sizes are scheduled regardless of the function under consideration, except for limited side-information such as smoothness or strong convexity parameters. With the knowledge of these two parameters, we show that any such algorithm attains an iteration complexity lower bound of Ω(\sqrtL/ε) for L-smooth convex functions, and \tildeΩ(\sqrtL/μ\ln(1/ε)) for L-smooth μ-strongly convex functions. These lower bounds are stronger than those in the traditional oracle model, as they hold independently of the dimension. To attain these, we abandon the oracle model in favor of a structure-based approach which builds upon a framework recently proposed in Arjevani et al. (2015). We further show that without knowing the strong convexity parameter, it is impossible to attain an iteration complexity better than \tildeΩ\sqrt(L/μ)\ln(1/ε). This res

ult is then used to formalize an observation regarding L-smooth convex functions , namely, that the iteration complexity of algorithms employing time-invariant s tep sizes must be at least $\Omega(L/\varepsilon)$.
****

Stochastic Variance Reduced Optimization for Nonconvex Sparse Learning
Xingguo Li, Tuo Zhao, Raman Arora, Han Liu, Jarvis Haupt
We propose a stochastic variance reduced optimization algorithm for solving a cl ass of large-scale nonconvex optimization problems with cardinality constraints, and provide sufficient conditions under which the proposed algorithm enjoys str ong linear convergence guarantees and optimal estimation accuracy in high dimens ions. Numerical experiments demonstrate the efficiency of our method in terms of both parameter estimation and computational performance.
****

Analysis of Variational Bayesian Factorizations for Sparse and Low-Rank Estimati on
David Wipf
Variational Bayesian (VB) approximations anchor a wide variety of probabilistic models, where tractable posterior inference is almost never possible. Typically based on the so-called VB mean-field approximation to the Kullback-Leibler diver gence, a posterior distribution is sought that factorizes across groups of laten t variables such that, with the distributions of all but one group of variables held fixed, an optimal closed-form distribution can be obtained for the remainin g group, with differing algorithms distinguished by how different variables are grouped and ultimately factored. This basic strategy is particularly attractive when estimating structured low-dimensional models of high-dimensional data, exem plified by the search for minimal rank and/or sparse approximations to observed data. To this end, VB models are frequently deployed across applications includi ng multi-task learning, robust PCA, subspace clustering, matrix completion, affi ne rank minimization, source localization, compressive sensing, and assorted com binations thereof. Perhaps surprisingly however, there exists almost no attendan t theoretical explanation for how various VB factorizations operate, and in whic h situations one may be preferable to another. We address this relative void by comparing arguably two of the most popular factorizations, one built upon Gaussi an scale mixture priors, the other bilinear Gaussian priors, both of which can f avor minimal rank or sparsity depending on the context. More specifically, by re expressing the respective VB objective functions, we weigh multiple factors rela ted to local minima avoidance, feature transformation invariance and correlation , and computational complexity to arrive at insightful conclusions useful in exp laining performance and deciding which VB flavor is advantageous. We also envisi on that the principles explored here are quite relevant to other structured inve rse problems where VB serves as a viable solution.
****

Fast k-means with accurate bounds
James Newling, Francois Fleuret
We propose a novel accelerated exact k-means algorithm, which outperforms the cu rrent state-of-the-art low-dimensional algorithm in 18 of 22 experiments, runnin g up to 3 times faster. We also propose a general improvement of existing state- of-the-art accelerated exact k-means algorithms through better estimates of the distance bounds used to reduce the number of distance calculations, obtaining sp eedups in 36 of 44 experiments, of up to 1.8 times. We have conducted experiment s with our own implementations of existing methods to ensure homogeneous evaluat ion of performance, and we show that our implementations perform as well or bett er than existing available implementations. Finally, we propose simplified varia nts of standard approaches and show that they are faster than their fully-fledge d counterparts in 59 of 62 experiments.
****

Boolean Matrix Factorization and Noisy Completion via Message Passing
Siamak Ravanbakhsh, Barnabas Poczos, Russell Greiner
Boolean matrix factorization and Boolean matrix completion from noisy observatio ns are desirable unsupervised data-analysis methods due to their interpretabilit

y, but hard to perform due to their NP-hardness. We treat these problems as maxi
mum a posteriori inference problems in a graphical model and present a message p
assing approach that scales linearly with the number of observations and factors
. Our empirical study demonstrates that message passing is able to recover low-r
ank Boolean matrices, in the boundaries of theoretically possible recovery and c
ompares favorably with state-of-the-art in real-world applications, such collabo
rative filtering with large-scale Boolean data.
****

Convolutional Rectifier Networks as Generalized Tensor Decompositions
Nadav Cohen, Amnon Shashua
Convolutional rectifier networks, i.e. convolutional neural networks with rectif
ied linear activation and max or average pooling, are the cornerstone of modern
deep learning. However, despite their wide use and success, our theoretical unde
rstanding of the expressive properties that drive these networks is partial at b
est. On the other hand, we have a much firmer grasp of these issues in the world
 of arithmetic circuits. Specifically, it is known that convolutional arithmetic
 circuits possess the property of "complete depth efficiency", meaning that besi
des a negligible set, all functions realizable by a deep network of polynomial s
ize, require exponential size in order to be realized (or approximated) by a sha
llow network. In this paper we describe a construction based on generalized tens
or decompositions, that transforms convolutional arithmetic circuits into convol
utional rectifier networks. We then use mathematical tools available from the wo
rld of arithmetic circuits to prove new results. First, we show that convolution
al rectifier networks are universal with max pooling but not with average poolin
g. Second, and more importantly, we show that depth efficiency is weaker with co
nvolutional rectifier networks than it is with convolutional arithmetic circuits
. This leads us to believe that developing effective methods for training convol
utional arithmetic circuits, thereby fulfilling their expressive potential, may
give rise to a deep learning architecture that is provably superior to convoluti
onal rectifier networks but has so far been overlooked by practitioners.
****

Low-rank Solutions of Linear Matrix Equations via Procrustes Flow
Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, Ben Recht
In this paper we study the problem of recovering a low-rank matrix from linear m
easurements. Our algorithm, which we call Procrustes Flow, starts from an initia
l estimate obtained by a thresholding scheme followed by gradient descent on a n
on-convex objective. We show that as long as the measurements obey a standard re
stricted isometry property, our algorithm converges to the unknown matrix at a g
eometric rate. In the case of Gaussian measurements, such convergence occurs for
 a n1 \times n2 matrix of rank r when the number of measurements exceeds a const
ant times (n1 + n2)r.
****

Anytime Exploration for Multi-armed Bandits using Confidence Information
Kwang-Sung Jun, Robert Nowak
We introduce anytime Explore-m, a pure exploration problem for multi-armed bandi
ts (MAB) that requires making a prediction of the top-m arms at every time step.
 Anytime Explore-m is more practical than fixed budget or fixed confidence formu
lations of the top-m problem, since many applications involve a finite, but unpr
edictable, budget. However, the development and analysis of anytime algorithms p
resent many challenges. We propose AT-LUCB (AnyTime Lower and Upper Confidence B
ound), the first nontrivial algorithm that provably solves anytime Explore-m. Ou
r analysis shows that the sample complexity of AT-LUCB is competitive to anytime
 variants of existing algorithms. Moreover, our empirical evaluation on AT-LUCB
shows that AT-LUCB performs as well as or better than state-of-the-art baseline
methods for anytime Explore-m.
****

Structured Prediction Energy Networks
David Belanger, Andrew McCallum
We introduce structured prediction energy networks (SPENs), a flexible framework
 for structured prediction. A deep architecture is used to define an energy func

tion of candidate labels, and then predictions are produced by using back-propag ation to iteratively optimize the energy with respect to the labels. This deep a rchitecture captures dependencies between labels that would lead to intractable graphical models, and performs structure learning by automatically learning disc riminative features of the structured output. One natural application of our tec hnique is multi-label classification, which traditionally has required strict pr ior assumptions about the interactions between labels to ensure tractable learni ng and prediction. We are able to apply SPENs to multi-label problems with subst antially larger label sets than previous applications of structured prediction, while modeling high-order interactions using minimal structural assumptions. Ove rall, deep learning provides remarkable tools for learning features of the input s to a prediction problem, and this work extends these techniques to learning fe atures of structured outputs. Our experiments provide impressive performance on a variety of benchmark multi-label classification tasks, demonstrate that our te chnique can be used to provide interpretable structure learning, and illuminate fundamental trade-offs between feed-forward and iterative structured prediction.
****

L1-regularized Neural Networks are Improperly Learnable in Polynomial Time
Yuchen Zhang, Jason D. Lee, Michael I. Jordan
We study the improper learning of multi-layer neural networks. Suppose that the neural network to be learned has k hidden layers and that the \ell_1-norm of the incoming weights of any neuron is bounded by L. We present a kernel-based metho d, such that with probability at least 1 - δ, it learns a predictor whose genera lization error is at most εworse than that of the neural network. The sample com plexity and the time complexity of the presented method are polynomial in the in put dimension and in (1/ε,\log(1/δ),F(k,L)), where F(k,L) is a function dependin g on (k,L) and on the activation function, independent of the number of neurons. The algorithm applies to both sigmoid-like activation functions and ReLU-like a ctivation functions. It implies that any sufficiently sparse neural network is l earnable in polynomial time.
****

Compressive Spectral Clustering
Nicolas Tremblay, Gilles Puy, Remi Gribonval, Pierre Vandergheynst
Spectral clustering has become a popular technique due to its high performance i n many contexts. It comprises three main steps: create a similarity graph betwee n N objects to cluster, compute the first k eigenvectors of its Laplacian matrix to define a feature vector for each object, and run k-means on these features t o separate objects into k classes. Each of these three steps becomes computation ally intensive for large N and/or k. We propose to speed up the last two steps b ased on recent results in the emerging field of graph signal processing: graph f iltering of random signals, and random sampling of bandlimited graph signals. We prove that our method, with a gain in computation time that can reach several o rders of magnitude, is in fact an approximation of spectral clustering, for whic h we are able to control the error. We test the performance of our method on art ificial and real-world network data.
****

Low-rank tensor completion: a Riemannian manifold preconditioning approach
Hiroyuki Kasai, Bamdev Mishra
We propose a novel Riemannian manifold preconditioning approach for the tensor c ompletion problem with rank constraint. A novel Riemannian metric or inner produ ct is proposed that exploits the least-squares structure of the cost function an d takes into account the structured symmetry that exists in Tucker decomposition . The specific metric allows to use the versatile framework of Riemannian optimi zation on quotient manifolds to develop preconditioned nonlinear conjugate gradi ent and stochastic gradient descent algorithms in batch and online setups, respe ctively. Concrete matrix representations of various optimization-related ingredi ents are listed. Numerical comparisons suggest that our proposed algorithms robu stly outperform state-of-the-art algorithms across different synthetic and real-world datasets.
****

Provable Non-convex Phase Retrieval with Outliers: Median TruncatedWirtinger Flow
Huishuai Zhang, Yuejie Chi, Yingbin Liang
Solving systems of quadratic equations is a central problem in machine learning and signal processing. One important example is phase retrieval, which aims to recover a signal from only magnitudes of its linear measurements. This paper focuses on the situation when the measurements are corrupted by arbitrary outliers, for which the recently developed non-convex gradient descent Wirtinger flow (WF) and truncated Wirtinger flow (TWF) algorithms likely fail. We develop a novel median-TWF algorithm that exploits robustness of sample median to resist arbitrary outliers in the initialization and the gradient update in each iteration. We show that such a non-convex algorithm provably recovers the signal from a near-optimal number of measurements composed of i.i.d. Gaussian entries, up to a logarithmic factor, even when a constant portion of the measurements are corrupted by arbitrary outliers. We further show that median-TWF is also robust when measurements are corrupted by both arbitrary outliers and bounded noise. Our analysis of performance guarantee is accomplished by development of non-trivial concentration measures of median-related quantities, which may be of independent interest. We further provide numerical experiments to demonstrate the effectiveness of the approach.
****
Estimating Maximum Expected Value through Gaussian Approximation
Carlo D’Eramo, Marcello Restelli, Alessandro Nuara
This paper is about the estimation of the maximum expected value of a set of independent random variables. The performance of several learning algorithms (e.g., Q-learning) is affected by the accuracy of such estimation. Unfortunately, no unbiased estimator exists. The usual approach of taking the maximum of the sample means leads to large overestimates that may significantly harm the performance of the learning algorithm. Recent works have shown that the cross validation estimator—which is negatively biased—outperforms the maximum estimator in many sequential decision-making scenarios. On the other hand, the relative performance of the two estimators is highly problem-dependent. In this paper, we propose a new estimator for the maximum expected value, based on a weighted average of the sample means, where the weights are computed using Gaussian approximations for the distributions of the sample means. We compare the proposed estimator with the other state-of-the-art methods both theoretically, by deriving upper bounds to the bias and the variance of the estimator, and empirically, by testing the performance on different sequential learning problems.
****
Representational Similarity Learning with Application to Brain Networks
Urvashi Oswal, Christopher Cox, Matthew Lambon-Ralph, Timothy Rogers, Robert Nowak
Representational Similarity Learning (RSL) aims to discover features that are important in representing (human-judged) similarities among objects. RSL can be posed as a sparsity-regularized multi-task regression problem. Standard methods, like group lasso, may not select important features if they are strongly correlated with others. To address this shortcoming we present a new regularizer for multitask regression called Group Ordered Weighted \ell_1 (GrOWL). Another key contribution of our paper is a novel application to fMRI brain imaging. Representational Similarity Analysis (RSA) is a tool for testing whether localized brain regions encode perceptual similarities. Using GrOWL, we propose a new approach called Network RSA that can discover arbitrarily structured brain networks (possibly widely distributed and non-local) that encode similarity information. We show, in theory and fMRI experiments, how GrOWL deals with strongly correlated covariates.
****
Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning
Yarin Gal, Zoubin Ghahramani
Deep learning tools have gained tremendous attention in applied machine learning

. However such tools for regression and classification do not capture model unce
rtainty. In comparison, Bayesian models offer a mathematically grounded framewor
k to reason about model uncertainty, but usually come with a prohibitive computa
tional cost. In this paper we develop a new theoretical framework casting dropou
t training in deep neural networks (NNs) as approximate Bayesian inference in de
ep Gaussian processes. A direct result of this theory gives us tools to model un
certainty with dropout NNs – extracting information from existing models that ha
s been thrown away so far. This mitigates the problem of representing uncertaint
y in deep learning without sacrificing either computational complexity or test a
ccuracy. We perform an extensive study of the properties of dropout's uncertaint
y. Various network architectures and non-linearities are assessed on tasks of re
gression and classification, using MNIST as an example. We show a considerable i
mprovement in predictive log-likelihood and RMSE compared to existing state-of-t
he-art methods, and finish by using dropout's uncertainty in deep reinforcement
learning.
****

Generative Adversarial Text to Image Synthesis
Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, Hong
lak Lee
Automatic synthesis of realistic images from text would be interesting and usefu
l, but current AI systems are still far from this goal. However, in recent years
 generic and powerful recurrent neural network architectures have been developed
 to learn discriminative text feature representations. Meanwhile, deep convoluti
onal generative adversarial networks (GANs) have begun to generate highly compel
ling images of specific categories such as faces, album covers, room interiors a
nd flowers. In this work, we develop a novel deep architecture and GAN formulati
on to effectively bridge these advances in text and image modeling, translating
visual concepts from characters to pixels. We demonstrate the capability of our
model to generate plausible images of birds and flowers from detailed text descr
iptions.
****

Dirichlet Process Mixture Model for Correcting Technical Variation in Single-Cel
l Gene Expression Data
Sandhya Prabhakaran, Elham Azizi, Ambrose Carr, Dana Pe'er
We introduce an iterative normalization and clustering method for single-cell ge
ne expression data. The emerging technology of single-cell RNA-seq gives access
to gene expression measurements for thousands of cells, allowing discovery and c
haracterization of cell types. However, the data is confounded by technical vari
ation emanating from experimental errors and cell type-specific biases. Current
approaches perform a global normalization prior to analyzing biological signals,
 which does not resolve missing data or variation dependent on latent cell types
. Our model is formulated as a hierarchical Bayesian mixture model with cell-spe
cific scalings that aid the iterative normalization and clustering of cells, tea
sing apart technical variation from biological signals. We demonstrate that this
 approach is superior to global normalization followed by clustering. We show id
entifiability and weak convergence guarantees of our method and present a scalab
le Gibbs inference algorithm. This method improves cluster inference in both syn
thetic and real single-cell data compared with previous methods, and allows easy
 interpretation and recovery of the underlying structure and cell types.
****

Improved SVRG for Non-Strongly-Convex or Sum-of-Non-Convex Objectives
Zeyuan Allen-Zhu, Yang Yuan
Many classical algorithms are found until several years later to outlive the con
fines in which they were conceived, and continue to be relevant in unforeseen se
ttings. In this paper, we show that SVRG is one such method: being originally de
signed for strongly convex objectives, it is also very robust in non-strongly co
nvex or sum-of-non-convex settings. More precisely, we provide new analysis to i
mprove the state-of-the-art running times in both settings by either applying SV
RG or its novel variant. Since non-strongly convex objectives include important
examples such as Lasso or logistic regression, and sum-of-non-convex objectives

include famous examples such as stochastic PCA and is even believed to be relate
d to training deep neural nets, our results also imply better performances in th
ese applications.
****

Sparse Parameter Recovery from Aggregated Data
Avradeep Bhowmik, Joydeep Ghosh, Oluwasanmi Koyejo
Data aggregation is becoming an increasingly common technique for sharing sensit
ive information, and for reducing data size when storage and/or communication co
sts are high. Aggregate quantities such as group-average are a form of semi-supe
rvision as they do not directly provide information of individual values, but de
spite their wide-spread use, prior literature on learning individual-level model
s from aggregated data is extremely limited. This paper investigates the effect
of data aggregation on parameter recovery for a sparse linear model, when known
results are no longer applicable. In particular, we consider a scenario where th
e data are collected into groups e.g. aggregated patient records, and first-orde
r empirical moments are available only at the group level. Despite this obfuscat
ion of individual data values, we can show that the true parameter is recoverabl
e with high probability using these aggregates when the collection of true group
 moments is an incoherent matrix, and the empirical moment estimates have been c
omputed from a sufficiently large number of samples. To the best of our knowledg
e, ours are the first results on structured parameter recovery using only aggreg
ated data. Experimental results on synthetic data are provided in support of the
se theoretical claims. We also show that parameter estimation from aggregated da
ta approaches the accuracy of parameter estimation obtainable from non-aggregate
d or “individual" samples, when applied to two real world healthcare application
s- predictive modeling of CMS Medicare reimbursement claims, and modeling of Tex
as State healthcare charges.
****

Deep Structured Energy Based Models for Anomaly Detection
Shuangfei Zhai, Yu Cheng, Weining Lu, Zhongfei Zhang
In this paper, we attack the anomaly detection problem by directly modeling the
data distribution with deep architectures. We hence propose deep structured ener
gy based models (DSEBMs), where the energy function is the output of a determini
stic deep neural network with structure. We develop novel model architectures to
 integrate EBMs with different types of data such as static data, sequential dat
a, and spatial data, and apply appropriate model architectures to adapt to the d
ata structure. Our training algorithm is built upon the recent development of sc
ore matching (Hyvarinen, 2005), which connects an EBM with a regularized autoenc
oder, eliminating the need for complicated sampling method. Statistically sound
decision criterion can be derived for anomaly detection purpose from the perspec
tive of the energy landscape of the data distribution. We investigate two decisi
on criteria for performing anomaly detection: the energy score and the reconstru
ction error. Extensive empirical studies on benchmark anomaly detection tasks de
monstrate that our proposed model consistently matches or outperforms all the co
mpeting methods.
****

Even Faster Accelerated Coordinate Descent Using Non-Uniform Sampling
Zeyuan Allen-Zhu, Zheng Qu, Peter Richtarik, Yang Yuan
Accelerated coordinate descent is widely used in optimization due to its cheap p
er-iteration cost and scalability to large-scale problems. Up to a primal-dual t
ransformation, it is also the same as accelerated stochastic gradient descent th
at is one of the central methods used in machine learning. In this paper, we imp
rove the best known running time of accelerated coordinate descent by a factor u
p to \sqrtn. Our improvement is based on a clean, novel non-uniform sampling tha
t selects each coordinate with a probability proportional to the square root of
its smoothness parameter. Our proof technique also deviates from the classical e
stimation sequence technique used in prior work. Our speed-up applies to importa
nt problems such as empirical risk minimization and solving linear systems, both
 in theory and in practice.
****

Unitary Evolution Recurrent Neural Networks
Martin Arjovsky, Amar Shah, Yoshua Bengio
Recurrent neural networks (RNNs) are notoriously difficult to train. When the ei genvalues of the hidden to hidden weight matrix deviate from absolute value 1, o ptimization becomes difficult due to the well studied issue of vanishing and exp loding gradients, especially when trying to learn long-term dependencies. To cir cumvent this problem, we propose a new architecture that learns a unitary weight matrix, with eigenvalues of absolute value exactly 1. The challenge we address is that of parametrizing unitary matrices in a way that does not require expensi ve computations (such as eigendecomposition) after each weight update. We constr uct an expressive unitary weight matrix by composing several structured matrices that act as building blocks with parameters to be learned. Optimization with th is parameterization becomes feasible only when considering hidden states in the complex domain. We demonstrate the potential of this architecture by achieving s tate of the art results in several hard tasks involving very long-term dependenc ies.
****

Markov Latent Feature Models
Aonan Zhang, John Paisley
We introduce Markov latent feature models (MLFM), a sparse latent feature model that arises naturally from a simple sequential construction. The key idea is to interpret each state of a sequential process as corresponding to a latent featur e, and the set of states visited between two null-state visits as picking out fe atures for an observation. We show that, given some natural constraints, we can represent this stochastic process as a mixture of recurrent Markov chains. In th is way we can perform correlated latent feature modeling for the sparse coding p roblem. We demonstrate two cases in which we define finite and infinite latent f eature models constructed from first-order Markov chains, and derive their assoc iated scalable inference algorithms. We show empirical results on a genome analy sis task and an image denoising task.
****

The Knowledge Gradient for Sequential Decision Making with Stochastic Binary Fee dbacks
Yingfei Wang, Chu Wang, Warren Powell
We consider the problem of sequentially making decisions that are rewarded by “s uccesses” and “failures” which can be predicted through an unknown relationship that depends on a partially controllable vector of attributes for each instance. The learner takes an active role in selecting samples from the instance pool. T he goal is to maximize the probability of success, either after the offline trai ning phase or minimizing regret in online learning. Our problem is motivated by real-world applications where observations are time consuming and/or expensive. With the adaptation of an online Bayesian linear classifier, we develop a knowle dge-gradient type policy to guide the experiment by maximizing the expected valu e of information of labeling each alternative, in order to reduce the number of expensive physical experiments. We provide a finite-time analysis of the estimat ed error and demonstrate the performance of the proposed algorithm on both synth etic problems and benchmark UCI datasets.
****

A Simple and Provable Algorithm for Sparse Diagonal CCA
Megasthenis Asteris, Anastasios Kyrillidis, Oluwasanmi Koyejo, Russell Poldrack
Given two sets of variables, derived from a common set of samples, sparse Canoni cal Correlation Analysis (CCA) seeks linear combinations of a small number of va riables in each set, such that the induced \emphcanonical variables are maximall y correlated. Sparse CCA is NP-hard. We propose a novel combinatorial algorithm for sparse diagonal CCA, \textiti.e., sparse CCA under the additional assumption that variables within each set are standardized and uncorrelated. Our algorithm operates on a low rank approximation of the input data and its computational co mplexity scales linearly with the number of input variables. It is simple to imp lement, and parallelizable. In contrast to most existing approaches, our algorit hm administers precise control on the sparsity of the extracted canonical vector

s, and comes with theoretical data-dependent global approximation guarantees, th at hinge on the spectrum of the input data. Finally, it can be straightforwardly adapted to other constrained variants of CCA enforcing structure beyond sparsit y. We empirically evaluate the proposed scheme and apply it on a real neuroimagi ng dataset to investigate associations between brain activity and behavior measu rements.
****

Quadratic Optimization with Orthogonality Constraints: Explicit Lojasiewicz Expo nent and Linear Convergence of Line-Search Methods
Huikang Liu, Weijie Wu, Anthony Man-Cho So
A fundamental class of matrix optimization problems that arise in many areas of science and engineering is that of quadratic optimization with orthogonality con straints. Such problems can be solved using line-search methods on the Stiefel m anifold, which are known to converge globally under mild conditions. To determin e the convergence rates of these methods, we give an explicit estimate of the ex ponent in a Lojasiewicz inequality for the (non-convex) set of critical points o f the aforementioned class of problems. This not only allows us to establish the linear convergence of a large class of line-search methods but also answers an important and intriguing problem in mathematical analysis and numerical optimiza tion. A key step in our proof is to establish a local error bound for the set of critical points, which may be of independent interest.
****

Normalization Propagation: A Parametric Technique for Removing Internal Covariat e Shift in Deep Networks
Devansh Arpit, Yingbo Zhou, Bhargava Kota, Venu Govindaraju
While the authors of Batch Normalization (BN) identify and address an important problem involved in training deep networks– \textitInternal Covariate Shift– the current solution has certain drawbacks. For instance, BN depends on batch stati stics for layerwise input normalization during training which makes the estimate s of mean and standard deviation of input (distribution) to hidden layers inaccu rate due to shifting parameter values (especially during initial training epochs). Another fundamental problem with BN is that it cannot be used with batch-size 1 during training. We address these drawbacks of BN by proposing a non-adapti ve normalization technique for removing covariate shift, that we call \textitNor malization Propagation. Our approach does not depend on batch statistics, but ra ther uses a data-independent parametric estimate of mean and standard-deviation in every layer thus being computationally faster compared with BN. We exploit th e observation that the pre-activation before Rectified Linear Units follow Gauss ian distribution in deep networks, and that once the first and second order stat istics of any given dataset are normalized, we can forward propagate this normal ization without the need for recalculating the approximate statistics for hidden layers.
****

Learning to Generate with Memory
Chongxuan Li, Jun Zhu, Bo Zhang
Memory units have been widely used to enrich the capabilities of deep networks o n capturing long-term dependencies in reasoning and prediction tasks, but little investigation exists on deep generative models (DGMs) which are good at inferri ng high-level invariant representations from unlabeled data. This paper presents a deep generative model with a possibly large external memory and an attention mechanism to capture the local detail information that is often lost in the bott om-up abstraction process in representation learning. By adopting a smooth atten tion model, the whole network is trained end-to-end by optimizing a variational bound of data likelihood via auto-encoding variational Bayesian methods, where a n asymmetric recognition network is learnt jointly to infer high-level invariant representations. The asymmetric architecture can reduce the competition between bottom-up invariant feature extraction and top-down generation of instance deta ils. Our experiments on several datasets demonstrate that memory can significant ly boost the performance of DGMs on various tasks, including density estimation, image generation, and missing value imputation, and DGMs with memory can achiev

e state-of-the-art quantitative results.
****
Learning End-to-end Video Classification with Rank-Pooling
Basura Fernando, Stephen Gould
We introduce a new model for representation learning and classification of video sequences. Our model is based on a convolutional neural network coupled with a novel temporal pooling layer. The temporal pooling layer relies on an inner-optimization problem to efficiently encode temporal semantics over arbitrarily long video clips into a fixed-length vector representation. Importantly, the representation and classification parameters of our model can be estimated jointly in an end-to-end manner by formulating learning as a bilevel optimization problem. Furthermore, the model can make use of any existing convolutional neural network architecture (e.g., AlexNet or VGG) without modification or introduction of additional parameters. We demonstrate our approach on action and activity recognition tasks.
****
Learning to Filter with Predictive State Inference Machines
Wen Sun, Arun Venkatraman, Byron Boots, J.Andrew Bagnell
Latent state space models are a fundamental and widely used tool for modeling dynamical systems. However, they are difficult to learn from data and learned models often lack performance guarantees on inference tasks such as filtering and prediction. In this work, we present the PREDICTIVE STATE INFERENCE MACHINE (PSIM), a data-driven method that considers the inference procedure on a dynamical system as a composition of predictors. The key idea is that rather than first learning a latent state space model, and then using the learned model for inference, PSIM directly learns predictors for inference in predictive state space. We provide theoretical guarantees for inference, in both realizable and agnostic settings, and showcase practical performance on a variety of simulated and real world robotics benchmarks.
****
A Subspace Learning Approach for High Dimensional Matrix Decomposition with Efficient Column/Row Sampling
Mostafa Rahmani, Geroge Atia
This paper presents a new randomized approach to high-dimensional low rank (LR) plus sparse matrix decomposition. For a data matrix D $\in$R^N_1 \times N_2, the complexity of conventional decomposition methods is O(N_1 N_2 r), which limits their usefulness in big data settings (r is the rank of the LR component). In addition, the existing randomized approaches rely for the most part on uniform random sampling, which may be inefficient for many real world data matrices. The proposed subspace learning based approach recovers the LR component using only a small subset of the columns/rows of data and reduces complexity to O(\max(N_1,N_2) r^2). Even when the columns/rows are sampled uniformly at random, the sufficient number of sampled columns/rows is shown to be roughly O(r μ), where $\mu$is the coherency parameter of the LR component. In addition, efficient sampling algorithms are proposed to address the problem of column/row sampling from structured data.
****
DCM Bandits: Learning to Rank with Multiple Clicks
Sumeet Katariya, Branislav Kveton, Csaba Szepesvari, Zheng Wen
A search engine recommends to the user a list of web pages. The user examines this list, from the first page to the last, and clicks on all attractive pages until the user is satisfied. This behavior of the user can be described by the dependent click model (DCM). We propose DCM bandits, an online learning variant of the DCM where the goal is to maximize the probability of recommending satisfactory items, such as web pages. The main challenge of our learning problem is that we do not observe which attractive item is satisfactory. We propose a computationally-efficient learning algorithm for solving our problem, dcmKL-UCB; derive gap-dependent upper bounds on its regret under reasonable assumptions; and also prove a matching lower bound up to logarithmic factors. We evaluate our algorithm on synthetic and real-world problems, and show that it performs well even when our model is misspecified. This work presents the first practical and regret-optim

al online algorithm for learning to rank with multiple clicks in a cascade-like click model.
****

Train faster, generalize better: Stability of stochastic gradient descent
Moritz Hardt, Ben Recht, Yoram Singer
We show that parametric models trained by a stochastic gradient method (SGM) with few iterations have vanishing generalization error. We prove our results by arguing that SGM is algorithmically stable in the sense of Bousquet and Elisseeff. Our analysis only employs elementary tools from convex and continuous optimization. We derive stability bounds for both convex and non-convex optimization under standard Lipschitz and smoothness assumptions. Applying our results to the convex case, we provide new insights for why multiple epochs of stochastic gradient methods generalize well in practice. In the non-convex case, we give a new interpretation of common practices in neural networks, and formally show that popular techniques for training large deep models are indeed stability-promoting. Our findings conceptually underscore the importance of reducing training time beyond its obvious benefit.
****

Copeland Dueling Bandit Problem: Regret Lower Bound, Optimal Algorithm, and Computationally Efficient Algorithm
Junpei Komiyama, Junya Honda, Hiroshi Nakagawa
We study the K-armed dueling bandit problem, a variation of the standard stochastic bandit problem where the feedback is limited to relative comparisons of a pair of arms. The hardness of recommending Copeland winners, the arms that beat the greatest number of other arms, is characterized by deriving an asymptotic regret bound. We propose Copeland Winners Deterministic Minimum Empirical Divergence (CW-RMED), an algorithm inspired by the DMED algorithm (Honda and Takemura, 2010), and derive an asymptotically optimal regret bound for it. However, it is not known whether the algorithm can be efficiently computed or not. To address this issue, we devise an efficient version (ECW-RMED) and derive its asymptotic regret bound. Experimental comparisons of dueling bandit algorithms show that ECW-RMED significantly outperforms existing ones.
****

Contextual Combinatorial Cascading Bandits
Shuai Li, Baoxiang Wang, Shengyu Zhang, Wei Chen
We propose the contextual combinatorial cascading bandits, a combinatorial online learning game, where at each time step a learning agent is given a set of contextual information, then selects a list of items, and observes stochastic outcomes of a prefix in the selected items by some stopping criterion. In online recommendation, the stopping criterion might be the first item a user selects; in network routing, the stopping criterion might be the first edge blocked in a path. We consider position discounts in the list order, so that the agent’s reward is discounted depending on the position where the stopping criterion is met. We design a UCB-type algorithm, C^3-UCB, for this problem, prove an n-step regret bound \tildeO(\sqrtn) in the general setting, and give finer analysis for two special cases. Our work generalizes existing studies in several directions, including contextual information, position discounts, and a more general cascading bandit model. Experiments on synthetic and real datasets demonstrate the advantage of involving contextual information and position discounts.
****

Conservative Bandits
Yifan Wu, Roshan Shariff, Tor Lattimore, Csaba Szepesvari
We study a novel multi-armed bandit problem that models the challenge faced by a company wishing to explore new strategies to maximize revenue whilst simultaneously maintaining their revenue above a fixed baseline, uniformly over time. While previous work addressed the problem under the weaker requirement of maintaining the revenue constraint only at a given fixed time in the future, the design of those algorithms makes them unsuitable under the more stringent constraints. We consider both the stochastic and the adversarial settings, where we propose natural yet novel strategies and analyze the price for maintaining the constraints.

Amongst other things, we prove both high probability and expectation bounds on the regret, while we also consider both the problem of maintaining the constraints with high probability or expectation. For the adversarial setting the price of maintaining the constraint appears to be higher, at least for the algorithm considered. A lower bound is given showing that the algorithm for the stochastic setting is almost optimal. Empirical results obtained in synthetic environments complement our theoretical findings.
****
Variance-Reduced and Projection-Free Stochastic Optimization
Elad Hazan, Haipeng Luo
The Frank-Wolfe optimization algorithm has recently regained popularity for machine learning applications due to its projection-free property and its ability to handle structured constraints. However, in the stochastic learning setting, it is still relatively understudied compared to the gradient descent counterpart. In this work, leveraging a recent variance reduction technique, we propose two stochastic Frank-Wolfe variants which substantially improve previous results in terms of the number of stochastic gradient evaluations needed to achieve 1-εaccuracy. For example, we improve from O(\frac1ε) to O(\ln\frac1ε) if the objective function is smooth and strongly convex, and from O(\frac1ε^2) to O(\frac1ε^1.5) if the objective function is smooth and Lipschitz. The theoretical improvement is also observed in experiments on real-world datasets for a multiclass classification application.
****
Factored Temporal Sigmoid Belief Networks for Sequence Learning
Jiaming Song, Zhe Gan, Lawrence Carin
Deep conditional generative models are developed to simultaneously learn the temporal dependencies of multiple sequences. The model is designed by introducing a three-way weight tensor to capture the multiplicative interactions between side information and sequences. The proposed model builds on the Temporal Sigmoid Belief Network (TSBN), a sequential stack of Sigmoid Belief Networks (SBNs). The transition matrices are further factored to reduce the number of parameters and improve generalization. When side information is not available, a general framework for semi-supervised learning based on the proposed model is constituted, allowing robust sequence classification. Experimental results show that the proposed approach achieves state-of-the-art predictive and classification performance on sequential data, and has the capacity to synthesize sequences, with controlled style transitioning and blending.
****
False Discovery Rate Control and Statistical Quality Assessment of Annotators in Crowdsourced Ranking
QianQian Xu, Jiechao Xiong, Xiaochun Cao, Yuan Yao
With the rapid growth of crowdsourcing platforms it has become easy and relatively inexpensive to collect a dataset labeled by multiple annotators in a short time. However due to the lack of control over the quality of the annotators, some abnormal annotators may be affected by position bias which can potentially degrade the quality of the final consensus labels. In this paper we introduce a statistical framework to model and detect annotator’s position bias in order to control the false discovery rate (FDR) without a prior knowledge on the amount of biased annotators–the expected fraction of false discoveries among all discoveries being not too high, in order to assure that most of the discoveries are indeed true and replicable. The key technical development relies on some new knockoff filters adapted to our problem and new algorithms based on the Inverse Scale Space dynamics whose discretization is potentially suitable for large scale crowdsourcing data analysis. Our studies are supported by experiments with both simulated examples and real-world data. The proposed framework provides us a useful tool for quantitatively studying annotator’s abnormal behavior in crowdsourcing.
****
Strongly-Typed Recurrent Neural Networks
David Balduzzi, Muhammad Ghifary
Recurrent neural networks are increasing popular models for sequential learning.

Unfortunately, although the most effective RNN architectures are perhaps excessively complicated, extensive searches have not found simpler alternatives. This paper imports ideas from physics and functional programming into RNN design to provide guiding principles. From physics, we introduce type constraints, analogous to the constraints that forbids adding meters to seconds. From functional programming, we require that strongly-typed architectures factorize into stateless learnware and state-dependent firmware, reducing the impact of side-effects. The features learned by strongly-typed nets have a simple semantic interpretation via a dynamic average-pooling on one-dimensional convolutions. We also show that strongly-typed gradients are better behaved than in classical architectures, and characterize the representational power of strongly-typed nets. Finally, experiments show that, despite being more constrained, strongly-typed architectures achieve lower training and comparable generalization error to classical architectures.

****

Distributed Clustering of Linear Bandits in Peer to Peer Networks
Nathan Korda, Balazs Szorenyi, Shuai Li

We provide two distributed confidence ball algorithms for solving linear bandit problems in peer to peer networks with limited communication capabilities. For the first, we assume that all the peers are solving the same linear bandit problem, and prove that our algorithm achieves the optimal asymptotic regret rate of any centralised algorithm that can instantly communicate information between the peers. For the second, we assume that there are clusters of peers solving the same bandit problem within each cluster, and we prove that our algorithm discovers these clusters, while achieving the optimal asymptotic regret rate within each one. Through experiments on several real-world datasets, we demonstrate the performance of proposed algorithms compared to the state-of-the-art.

****

Collapsed Variational Inference for Sum-Product Networks
Han Zhao, Tameem Adel, Geoff Gordon, Brandon Amos

Sum-Product Networks (SPNs) are probabilistic inference machines that admit exact inference in linear time in the size of the network. Existing parameter learning approaches for SPNs are largely based on the maximum likelihood principle and are subject to overfitting compared to more Bayesian approaches. Exact Bayesian posterior inference for SPNs is computationally intractable. Even approximation techniques such as standard variational inference and posterior sampling for SPNs are computationally infeasible even for networks of moderate size due to the large number of local latent variables per instance. In this work, we propose a novel deterministic collapsed variational inference algorithm for SPNs that is computationally efficient, easy to implement and at the same time allows us to incorporate prior information into the optimization formulation. Extensive experiments show a significant improvement in accuracy compared with a maximum likelihood based approach.

****

On the Analysis of Complex Backup Strategies in Monte Carlo Tree Search
Piyush Khandelwal, Elad Liebman, Scott Niekum, Peter Stone

Over the past decade, Monte Carlo Tree Search (MCTS) and specifically Upper Confidence Bound in Trees (UCT) have proven to be quite effective in large probabilistic planning domains. In this paper, we focus on how values are backpropagated in the MCTS tree, and apply complex return strategies from the Reinforcement Learning (RL) literature to MCTS, producing 4 new MCTS variants. We demonstrate that in some probabilistic planning benchmarks from the International Planning Competition (IPC), selecting a MCTS variant with a backup strategy different from Monte Carlo averaging can lead to substantially better results. We also propose a hypothesis for why different backup strategies lead to different performance in particular environments, and manipulate a carefully structured grid-world domain to provide empirical evidence supporting our hypothesis.

****

Benchmarking Deep Reinforcement Learning for Continuous Control
Yan Duan, Xi Chen, Rein Houthooft, John Schulman, Pieter Abbeel

Recently, researchers have made significant progress combining the advances in d eep learning for learning feature representations with reinforcement learning. S ome notable examples include training agents to play Atari games based on raw pi xel data and to acquire advanced manipulation skills using raw sensory inputs. H owever, it has been difficult to quantify progress in the domain of continuous c ontrol due to the lack of a commonly adopted benchmark. In this work, we present a benchmark suite of continuous control tasks, including classic tasks like car t-pole swing-up, tasks with very high state and action dimensionality such as 3D humanoid locomotion, tasks with partial observations, and tasks with hierarchic al structure. We report novel findings based on the systematic evaluation of a r ange of implemented reinforcement learning algorithms. Both the benchmark and re ference implementations are released at https://github.com/rllab/rllab in order to facilitate experimental reproducibility and to encourage adoption by other re searchers.
****

K-Means Clustering with Distributed Dimensions
Hu Ding, Yu Liu, Lingxiao Huang, Jian Li
Distributed clustering has attracted significant attention in recent years. In t his paper, we study the k-means problem in the distributed dimension setting, wh ere the dimensions of the data are partitioned across multiple machines. We prov ide new approximation algorithms, which incur low communication costs and achiev e constant approximation ratios. The communication complexity of our algorithms significantly improve on existing algorithms. We also provide the first communic ation lower bound, which nearly matches our upper bound in a certain range of pa rameter setting. Our experimental results show that our algorithms outperform ex isting algorithms on real data-sets in the distributed dimension setting.
****

Texture Networks: Feed-forward Synthesis of Textures and Stylized Images
Dmitry Ulyanov, Vadim Lebedev, Andrea, Victor Lempitsky
Gatys et al. recently demonstrated that deep networks can generate beautiful tex tures and stylized images from a single texture example. However, their methods requires a slow and memory-consuming optimization process. We propose here an al ternative approach that moves the computational burden to a learning stage. Give n a single example of a texture, our approach trains compact feed-forward convol utional networks to generate multiple samples of the same texture of arbitrary s ize and to transfer artistic style from a given image to any other image. The re sulting networks are remarkably light-weight and can generate textures of qualit y comparable to Gatys et al., but hundreds of times faster. More generally, our approach highlights the power and flexibility of generative feed-forward models trained with complex and expressive loss functions.
****

Fast Constrained Submodular Maximization: Personalized Data Summarization
Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, Amin Karbasi
Can we summarize multi-category data based on user preferences in a scalable man ner? Many utility functions used for data summarization satisfy submodularity, a natural diminishing returns property. We cast personalized data summarization a s an instance of a general submodular maximization problem subject to multiple c onstraints. We develop the first practical and FAst coNsTrained submOdular Maxim ization algorithm, FANTOM, with strong theoretical guarantees. FANTOM maximizes a submodular function (not necessarily monotone) subject to intersection of a p-system and l knapsacks constrains. It achieves a $(1 + \varepsilon)(p + 1)(2p + 2l + 1)/p$ a pproximation guarantee with only $O(nrp \log(n)/\varepsilon)$ query complexity (n and r indic ate the size of the ground set and the size of the largest feasible solution, re spectively). We then show how we can use FANTOM for personalized data summarizat ion. In particular, a p-system can model different aspects of data, such as cate gories or time stamps, from which the users choose. In addition, knapsacks encod e users' constraints including budget or time. In our set of experiments, we con sider several concrete applications: movie recommendation over 11K movies, perso nalized image summarization with 10K images, and revenue maximization on the You Tube social networks with 5000 communities. We observe that FANTOM constantly pr

ovides the highest utility against all the baselines.
****

On the Statistical Limits of Convex Relaxations
Zhaoran Wang, Quanquan Gu, Han Liu
Many high dimensional sparse learning problems are formulated as nonconvex optim ization. A popular approach to solve these nonconvex optimization problems is th rough convex relaxations such as linear and semidefinite programming. In this pa per, we study the statistical limits of convex relaxations. Particularly, we con sider two problems: Mean estimation for sparse principal submatrix and edge prob ability estimation for stochastic block model. We exploit the sum-of-squares rel axation hierarchy to sharply characterize the limits of a broad class of convex relaxations. Our result shows statistical optimality needs to be compromised for achieving computational tractability using convex relaxations. Compared with ex isting results on computational lower bounds for statistical problems, which con sider general polynomial-time algorithms and rely on computational hardness hypo theses on problems like planted clique detection, our theory focuses on a broad class of convex relaxations and does not rely on unproven hypotheses.
****

Ask Me Anything: Dynamic Memory Networks for Natural Language Processing
Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gul rajani, Victor Zhong, Romain Paulus, Richard Socher
Most tasks in natural language processing can be cast into question answering (Q A) problems over language input. We introduce the dynamic memory network (DMN), a neural network architecture which processes input sequences and questions, for ms episodic memories, and generates relevant answers. Questions trigger an itera tive attention process which allows the model to condition its attention on the inputs and the result of previous iterations. These results are then reasoned ov er in a hierarchical recurrent sequence model to generate answers. The DMN can b e trained end-to-end and obtains state-of-the-art results on several types of ta sks and datasets: question answering (Facebook’s bAbI dataset), text classificat ion for sentiment analysis (Stanford Sentiment Treebank) and sequence modeling f or part-of-speech tagging (WSJ-PTB). The training for these different tasks reli es exclusively on trained word vector representations and input-question-answer triplets.
****

Gossip Dual Averaging for Decentralized Optimization of Pairwise Functions
Igor Colin, Aurelien Bellet, Joseph Salmon, Stéphan Clémençon
In decentralized networks (of sensors, connected objects, etc.), there is an imp ortant need for efficient algorithms to optimize a global cost function, for ins tance to learn a global model from the local data collected by each computing un it. In this paper, we address the problem of decentralized minimization of pairw ise functions of the data points, where these points are distributed over the no des of a graph defining the communication topology of the network. This general problem finds applications in ranking, distance metric learning and graph infere nce, among others. We propose new gossip algorithms based on dual averaging whic h aims at solving such problems both in synchronous and asynchronous settings. T he proposed framework is flexible enough to deal with constrained and regularize d variants of the optimization problem. Our theoretical analysis reveals that th e proposed algorithms preserve the convergence rate of centralized dual averagin g up to an additive bias term. We present numerical simulations on Area Under th e ROC Curve (AUC) maximization and metric learning problems which illustrate the practical interest of our approach.
****

Solving Ridge Regression using Sketched Preconditioned SVRG
Alon Gonen, Francesco Orabona, Shai Shalev-Shwartz
We develop a novel preconditioning method for ridge regression, based on recent linear sketching methods. By equipping Stochastic Variance Reduced Gradient (SVR G) with this preconditioning process, we obtain a significant speed-up relative to fast stochastic methods such as SVRG, SDCA and SAG.
****

Cumulative Prospect Theory Meets Reinforcement Learning: Prediction and Control
Prashanth L.A., Cheng Jie, Michael Fu, Steve Marcus, Csaba Szepesvari
Cumulative prospect theory (CPT) is known to model human decisions well, with su
bstantial empirical evidence supporting this claim. CPT works by distorting prob
abilities and is more general than the classic expected utility and coherent ris
k measures. We bring this idea to a risk-sensitive reinforcement learning (RL) s
etting and design algorithms for both estimation and control. The RL setting pre
sents two particular challenges when CPT is applied: estimating the CPT objectiv
e requires estimations of the entire distribution of the value function and find
ing a randomized optimal policy. The estimation scheme that we propose uses the
empirical distribution to estimate the CPT-value of a random variable. We then u
se this scheme in the inner loop of a CPT-value optimization procedure that is b
ased on the well-known simulation optimization idea of simultaneous perturbation
 stochastic approximation (SPSA). We provide theoretical convergence guarantees
for all the proposed algorithms and also empirically demonstrate the usefulness
of our algorithms.
****

Estimating Accuracy from Unlabeled Data: A Bayesian Approach
Emmanouil Antonios Platanios, Avinava Dubey, Tom Mitchell
We consider the question of how unlabeled data can be used to estimate the true
accuracy of learned classifiers, and the related question of how outputs from se
veral classifiers performing the same task can be combined based on their estima
ted accuracies. To answer these questions, we first present a simple graphical m
odel that performs well in practice. We then provide two nonparametric extension
s to it that improve its performance. Experiments on two real-world data sets pr
oduce accuracy estimates within a few percent of the true accuracy, using solely
 unlabeled data. Our models also outperform existing state-of-the-art solutions
in both estimating accuracies, and combining multiple classifier outputs.
****

Non-negative Matrix Factorization under Heavy Noise
Chiranjib Bhattacharya, Navin Goyal, Ravindran Kannan, Jagdeep Pani
The Noisy Non-negative Matrix factorization (NMF) is: given a data matrix A (d x
 n), find non-negative matrices B;C (d x k, k x n respy.) so that A = BC +N, whe
re N is a noise matrix. Existing polynomial time algorithms with proven error gu
arantees require EACH column N_.j to have l1 norm much smaller than ||(BC)_.j ||
_1, which could be very restrictive. In important applications of NMF such as To
pic Modeling as well as theoretical noise models (e.g. Gaussian with high sigma)
, almost EVERY column of N_.j violates this condition. We introduce the heavy no
ise model which only requires the average noise over large subsets of columns to
 be small. We initiate a study of Noisy NMF under the heavy noise model. We show
 that our noise model subsumes noise models of theoretical and practical interes
t (for e.g. Gaussian noise of maximum possible sigma). We then devise an algorit
hm TSVDNMF which under certain assumptions on B,C, solves the problem under heav
y noise. Our error guarantees match those of previous algorithms. Our running ti
me of O(k.(d+n)^2) is substantially better than the O(d.n^3) for the previous be
st. Our assumption on B is weaker than the “Separability” assumption made by all
 previous results. We provide empirical justification for our assumptions on C.
We also provide the first proof of identifiability (uniqueness of B) for noisy N
MF which is not based on separability and does not use hard to check geometric c
onditions. Our algorithm outperforms earlier polynomial time algorithms both in
time and error, particularly in the presence of high noise.
****

Extreme F-measure Maximization using Sparse Probability Estimates
Kalina Jasinska, Krzysztof Dembczynski, Robert Busa-Fekete, Karlson Pfannschmidt
, Timo Klerx, Eyke Hullermeier
We consider the problem of (macro) F-measure maximization in the context of extr
eme multi-label classification (XMLC), i.e., multi-label classification with ext
remely large label spaces. We investigate several approaches based on recent res
ults on the maximization of complex performance measures in binary classificatio
n. According to these results, the F-measure can be maximized by properly thresh

olding conditional class probability estimates. We show that a naive adaptation of this approach can be very costly for XMLC and propose to solve the problem by classifiers that efficiently deliver sparse probability estimates (SPEs), that is, probability estimates restricted to the most probable labels. Empirical results provide evidence for the strong practical performance of this approach.

****

Auxiliary Deep Generative Models

Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, Ole Winther

Deep generative models parameterized by neural networks have recently achieved state-of-the-art performance in unsupervised and semi-supervised learning. We extend deep generative models with auxiliary variables which improves the variational approximation. The auxiliary variables leave the generative model unchanged but make the variational distribution more expressive. Inspired by the structure of the auxiliary variable we also propose a model with two stochastic layers and skip connections. Our findings suggest that more expressive and properly specified deep generative models converge faster with better results. We show state-of-the-art performance within semi-supervised learning on MNIST, SVHN and NORB datasets.

****

Importance Sampling Tree for Large-scale Empirical Expectation

Olivier Canevet, Cijo Jose, Francois Fleuret

We propose a tree-based procedure inspired by the Monte-Carlo Tree Search that dynamically modulates an importance-based sampling to prioritize computation, while getting unbiased estimates of weighted sums. We apply this generic method to learning on very large training sets, and to the evaluation of large-scale SVMs. The core idea is to reformulate the estimation of a score - whether a loss or a prediction estimate - as an empirical expectation, and to use such a tree whose leaves carry the samples to focus efforts over the problematic "heavy weight" ones. We illustrate the potential of this approach on three problems: to improve Adaboost and a multi-layer perceptron on 2D synthetic tasks with several million points, to train a large-scale convolution network on several millions deformations of the CIFAR data-set, and to compute the response of a SVM with several hundreds of thousands of support vectors. In each case, we show how it either cuts down computation by more than one order of magnitude and/or allows to get better loss estimates.

****

Starting Small - Learning with Adaptive Sample Sizes

Hadi Daneshmand, Aurelien Lucchi, Thomas Hofmann

For many machine learning problems, data is abundant and it may be prohibitive to make multiple passes through the full training set. In this context, we investigate strategies for dynamically increasing the effective sample size, when using iterative methods such as stochastic gradient descent. Our interest is motivated by the rise of variance-reduced methods, which achieve linear convergence rates that scale favorably for smaller sample sizes. Exploiting this feature, we show - theoretically and empirically - how to obtain significant speed-ups with a novel algorithm that reaches statistical accuracy on an n-sample in 2n, instead of n log n steps.

****

Deep Gaussian Processes for Regression using Approximate Expectation Propagation

Thang Bui, Daniel Hernandez-Lobato, Jose Hernandez-Lobato, Yingzhen Li, Richard Turner

Deep Gaussian processes (DGPs) are multi-layer hierarchical generalisations of Gaussian processes (GPs) and are formally equivalent to neural networks with multiple, infinitely wide hidden layers. DGPs are nonparametric probabilistic models and as such are arguably more flexible, have a greater capacity to generalise, and provide better calibrated uncertainty estimates than alternative deep models. This paper develops a new approximate Bayesian learning scheme that enables DGPs to be applied to a range of medium to large scale regression problems for the first time. The new method uses an approximate Expectation Propagation procedure and a novel and efficient extension of the probabilistic backpropagation algor

ithm for learning. We evaluate the new method for non-linear regression on eleve n real-world datasets, showing that it always outperforms GP regression and is a lmost always better than state-of-the-art deterministic and sampling-based appro ximate inference methods for Bayesian neural networks. As a by-product, this wor k provides a comprehensive analysis of six approximate Bayesian methods for trai ning neural networks.
****

DR-ABC: Approximate Bayesian Computation with Kernel-Based Distribution Regressi on

Jovana Mitrovic, Dino Sejdinovic, Yee-Whye Teh

Performing exact posterior inference in complex generative models is often diffi cult or impossible due to an expensive to evaluate or intractable likelihood fun ction. Approximate Bayesian computation (ABC) is an inference framework that con structs an approximation to the true likelihood based on the similarity between the observed and simulated data as measured by a predefined set of summary stati stics. Although the choice of informative problem-specific summary statistics cr ucially influences the quality of the likelihood approximation and hence also th e quality of the posterior sample in ABC, there are only few principled general- purpose approaches to the selection or construction of such summary statistics. In this paper, we develop a novel framework for solving this problem. We model t he functional relationship between the data and the optimal choice (with respect to a loss function) of summary statistics using kernel-based distribution regre ssion. Furthermore, we extend our approach to incorporate kernel-based regressio n from conditional distributions, thus appropriately taking into account the spe cific structure of the posited generative model. We show that our approach can b e implemented in a computationally and statistically efficient way using the ran dom Fourier features framework for large-scale kernel learning. In addition to t hat, our framework outperforms related methods by a large margin on toy and real -world data, including hierarchical and time series models.
****

Predictive Entropy Search for Multi-objective Bayesian Optimization

Daniel Hernandez-Lobato, Jose Hernandez-Lobato, Amar Shah, Ryan Adams

We present \small PESMO, a Bayesian method for identifying the Pareto set of mul ti-objective optimization problems, when the functions are expensive to evaluate . \small PESMO chooses the evaluation points to maximally reduce the entropy of the posterior distribution over the Pareto set. The \small PESMO acquisition fun ction is decomposed as a sum of objective-specific acquisition functions, which makes it possible to use the algorithm in \emphdecoupled scenarios in which the objectives can be evaluated separately and perhaps with different costs. This de coupling capability is useful to identify difficult objectives that require more evaluations. \small PESMO also offers gains in efficiency, as its cost scales l inearly with the number of objectives, in comparison to the exponential cost of other methods. We compare \small PESMO with other methods on synthetic and real- world problems. The results show that \small PESMO produces better recommendatio ns with a smaller number of evaluations, and that a decoupled evaluation can lea d to improvements in performance, particularly when the number of objectives is large.
****

Rich Component Analysis

Rong Ge, James Zou

In many settings, we have multiple data sets (also called views) that capture di fferent and overlapping aspects of the same phenomenon. We are often interested in finding patterns that are unique to one or to a subset of the views. For exam ple, we might have one set of molecular observations and one set of physiologica l observations on the same group of individuals, and we want to quantify molecul ar patterns that are uncorrelated with physiology. Despite being a common proble m, this is highly challenging when the correlations come from complex distributi ons. In this paper, we develop the general framework of Rich Component Analysis (RCA) to model settings where the observations from different views are driven b y different sets of latent components, and each component can be a complex, high

-dimensional distribution. We introduce algorithms based on cumulant extraction that provably learn each of the components without having to model the other components. We show how to integrate RCA with stochastic gradient descent into a meta-algorithm for learning general models, and demonstrate substantial improvement in accuracy on several synthetic and real datasets in both supervised and unsupervised tasks. Our method makes it possible to learn latent variable models when we don’t have samples from the true model but only samples after complex perturbations.
****

Black-Box Alpha Divergence Minimization
Jose Hernandez-Lobato, Yingzhen Li, Mark Rowland, Thang Bui, Daniel Hernandez-Lobato, Richard Turner
Black-box alpha (BB-α) is a new approximate inference method based on the minimization of α-divergences. BB-αscales to large datasets because it can be implemented using stochastic gradient descent. BB-αcan be applied to complex probabilistic models with little effort since it only requires as input the likelihood function and its gradients. These gradients can be easily obtained using automatic differentiation. By changing the divergence parameter α, the method is able to interpolate between variational Bayes (VB) (α→0) and an algorithm similar to expectation propagation (EP) (α= 1). Experiments on probit regression and neural network regression and classification problems show that BB-αwith non-standard settings of α, such as α= 0.5, usually produces better predictions than with α→0 (VB) or α= 1 (EP).
****

One-Shot Generalization in Deep Generative Models
Danilo Rezende, Shakir, Ivo Danihelka, Karol Gregor, Daan Wierstra
Humans have an impressive ability to reason about new concepts and experiences from just a single example. In particular, humans have an ability for one-shot generalization: an ability to encounter a new concept, understand its structure, and then be able to generate compelling alternative variations of the concept. We develop machine learning systems with this important capacity by developing new deep generative models, models that combine the representational power of deep learning with the inferential power of Bayesian reasoning. We develop a class of sequential generative models that are built on the principles of feedback and attention. These two characteristics lead to generative models that are among the state-of-the art in density estimation and image generation. We demonstrate the one-shot generalization ability of our models using three tasks: unconditional sampling, generating new exemplars of a given concept, and generating new exemplars of a family of concepts. In all cases our models are able to generate compelling and diverse samples—having seen new examples just once—providing an important class of general-purpose models for one-shot machine learning.
****

Optimal Classification with Multivariate Losses
Nagarajan Natarajan, Oluwasanmi Koyejo, Pradeep Ravikumar, Inderjit Dhillon
Multivariate loss functions are extensively employed in several prediction tasks arising in Information Retrieval. Often, the goal in the tasks is to minimize expected loss when retrieving relevant items from a presented set of items, where the expectation is with respect to the joint distribution over item sets. Our key result is that for most multivariate losses, the expected loss is provably optimized by sorting the items by the conditional probability of label being positive and then selecting top k items. Such a result was previously known only for the F-measure. Leveraging on the optimality characterization, we give an algorithm for estimating optimal predictions in practice with runtime quadratic in size of item sets for many losses. We provide empirical results on benchmark datasets, comparing the proposed algorithm to state-of-the-art methods for optimizing multivariate losses.
****

A ranking approach to global optimization
Cedric Malherbe, Emile Contal, Nicolas Vayatis
We consider the problem of maximizing an unknown function f over a compact and c

onvex set using as few observations f(x) as possible. We observe that the optimi zation of the function f essentially relies on learning the induced bipartite ra nking rule of f. Based on this idea, we relate global optimization to bipartite ranking which allows to address problems with high dimensional input space, as w ell as cases of functions with weak regularity properties. The paper introduces novel meta-algorithms for global optimization which rely on the choice of any bi partite ranking method. Theoretical properties are provided as well as convergen ce guarantees and equivalences between various optimization methods are obtained as a by-product. Eventually, numerical evidence is given to show that the main algorithm of the paper which adapts empirically to the underlying ranking struct ure essentially outperforms existing state-of-the-art global optimization algori thms in typical benchmarks.

****

Parallel and Distributed Block-Coordinate Frank-Wolfe Algorithms

Yu-Xiang Wang, Veeranjaneyulu Sadhanala, Wei Dai, Willie Neiswanger, Suvrit Sra, Eric Xing

We study parallel and distributed Frank-Wolfe algorithms; the former on shared m emory machines with mini-batching, and the latter in a delayed update framework. In both cases, we perform computations asynchronously whenever possible. We ass ume block-separable constraints as in Block-Coordinate Frank-Wolfe (BCFW) method (Lacoste et. al., 2013) , but our analysis subsumes BCFW and reveals problem-de pendent quantities that govern the speedups of our methods over BCFW. A notable feature of our algorithms is that they do not depend on worst-case bounded delay s, but only (mildly) on **expected** delays, making them robust to stragglers an d faulty worker threads. We present experiments on structural SVM and Group Fuse d Lasso, and observe significant speedups over competing state-of-the-art (and s ynchronous) methods.

****

Autoencoding beyond pixels using a learned similarity metric

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, Ole Winther

We present an autoencoder that leverages learned representations to better measu re similarities in data space. By combining a variational autoencoder (VAE) with a generative adversarial network (GAN) we can use learned feature representatio ns in the GAN discriminator as basis for the VAE reconstruction objective. There by, we replace element-wise errors with feature-wise errors to better capture th e data distribution while offering invariance towards e.g. translation. We apply our method to images of faces and show that it outperforms VAEs with element-wi se similarity measures in terms of visual fidelity. Moreover, we show that the m ethod learns an embedding in which high-level abstract visual features (e.g. wea ring glasses) can be modified using simple arithmetic.

****

Ensuring Rapid Mixing and Low Bias for Asynchronous Gibbs Sampling

Christopher De Sa, Chris Re, Kunle Olukotun

Gibbs sampling is a Markov chain Monte Carlo technique commonly used for estimat ing marginal distributions. To speed up Gibbs sampling, there has recently been interest in parallelizing it by executing asynchronously. While empirical result s suggest that many models can be efficiently sampled asynchronously, traditiona l Markov chain analysis does not apply to the asynchronous case, and thus asynch ronous Gibbs sampling is poorly understood. In this paper, we derive a better un derstanding of the two main challenges of asynchronous Gibbs: bias and mixing ti me. We show experimentally that our theoretical results match practical outcomes .

****

Simultaneous Safe Screening of Features and Samples in Doubly Sparse Modeling

Atsushi Shibagaki, Masayuki Karasuyama, Kohei Hatano, Ichiro Takeuchi

The problem of learning a sparse model is conceptually interpreted as the proces s of identifying active features/samples and then optimizing the model over them . Recently introduced safe screening allows us to identify a part of non-active features/samples. So far, safe screening has been individually studied either fo r feature screening or for sample screening. In this paper, we introduce a new a

pproach for safely screening features and samples simultaneously by alternativel
y iterating feature and sample screening steps. A significant advantage of consi
dering them simultaneously rather than individually is that they have a synergy
effect in the sense that the results of the previous safe feature screening can
be exploited for improving the next safe sample screening performances, and vice
-versa. We first theoretically investigate the synergy effect, and then illustra
te the practical advantage through intensive numerical experiments for problems
with large numbers of features and samples.
****

Anytime optimal algorithms in stochastic multi-armed bandits
Rémy Degenne, Vianney Perchet
We introduce an anytime algorithm for stochastic multi-armed bandit with optimal
 distribution free and distribution dependent bounds (for a specific family of p
arameters). The performances of this algorithm (as well as another one motivated
 by the conjectured optimal bound) are evaluated empirically. A similar analysis
 is provided with full information, to serve as a benchmark.
****

Bounded Off-Policy Evaluation with Missing Data for Course Recommendation and Cu
rriculum Design
William Hoiles, Mihaela Schaar
Successfully recommending personalized course schedules is a difficult problem g
iven the diversity of students knowledge, learning behaviour, and goals. This pa
per presents personalized course recommendation and curriculum design algorithms
 that exploit logged student data. The algorithms are based on the regression es
timator for contextual multi-armed bandits with a penalized variance term. Guara
ntees on the predictive performance of the algorithms are provided using empiric
al Bernstein bounds. We also provide guidelines for including expert domain know
ledge into the recommendations. Using undergraduate engineering logged data from
 a post-secondary institution we illustrate the performance of these algorithms.
****

On collapsed representation of hierarchical Completely Random Measures
Gaurav Pandey, Ambedkar Dukkipati
The aim of the paper is to provide an exact approach for generating a Poisson pr
ocess sampled from a hierarchical CRM, without having to instantiate the infinit
ely many atoms of the random measures. We use completely random measures (CRM) a
nd hierarchical CRM to define a prior for Poisson processes. We derive the margi
nal distribution of the resultant point process, when the underlying CRM is marg
inalized out. Using well known properties unique to Poisson processes, we were a
ble to derive an exact approach for instantiating a Poisson process with a hiera
rchical CRM prior. Furthermore, we derive Gibbs sampling strategies for hierarch
ical CRM models based on Chinese restaurant franchise sampling scheme. As an exa
mple, we present the sum of generalized gamma process (SGGP), and show its appli
cation in topic-modelling. We show that one can determine the power-law behaviou
r of the topics and words in a Bayesian fashion, by defining a prior on the para
meters of SGGP.
****

From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classific
ation
Andre Martins, Ramon Astudillo
We propose sparsemax, a new activation function similar to the traditional softm
ax, but able to output sparse probabilities. After deriving its properties, we s
how how its Jacobian can be efficiently computed, enabling its use in a network
trained with backpropagation. Then, we propose a new smooth and convex loss func
tion which is the sparsemax analogue of the logistic loss. We reveal an unexpect
ed connection between this new loss and the Huber classification loss. We obtain
 promising empirical results in multi-label classification problems and in atten
tion-based neural networks for natural language inference. For the latter, we ac
hieve a similar performance as the traditional softmax, but with a selective, mo
re compact, attention focus.
****

Black-box Optimization with a Politician
Sebastien Bubeck, Yin Tat Lee
We propose a new framework for black-box convex optimization which is well-suite d for situations where gradient computations are expensive. We derive a new meth od for this framework which leverages several concepts from convex optimization, from standard first-order methods (e.g. gradient descent or quasi-Newton method s) to analytical centers (i.e. minimizers of self-concordant barriers). We demon strate empirically that our new technique compares favorably with state of the a rt algorithms (such as BFGS).
****
Gaussian process nonparametric tensor estimator and its minimax optimality
Heishiro Kanagawa, Taiji Suzuki, Hayato Kobayashi, Nobuyuki Shimizu, Yukihiro Ta gami
We investigate the statistical efficiency of a nonparametric Gaussian process me thod for a nonlinear tensor estimation problem. Low-rank tensor estimation has b een used as a method to learn higher order relations among several data sources in a wide range of applications, such as multi-task learning, recommendation sys tems, and spatiotemporal analysis. We consider a general setting where a common linear tensor learning is extended to a nonlinear learning problem in reproducin g kernel Hilbert space and propose a nonparametric Bayesian method based on the Gaussian process method. We prove its statistical convergence rate without assum ing any strong convexity, such as restricted strong convexity. Remarkably, it is shown that our convergence rate achieves the minimax optimal rate. We apply our proposed method to multi-task learning and show that our method significantly o utperforms existing methods through numerical experiments on real-world data set s.
****
No-Regret Algorithms for Heavy-Tailed Linear Bandits
Andres Munoz Medina, Scott Yang
We analyze the problem of linear bandits under heavy tailed noise. Most of of th e work on linear bandits has been based on the assumption of bounded or sub-Gaus sian noise. However, this assumption is often violated in common scenarios such as financial markets. We present two algorithms to tackle this problem: one base d on dynamic truncation and one based on a median of means estimator. We show th at, when the noise admits admits only a 1 + εmoment, these algorithms are still able to achieve regret in \widetildeO(T^\frac2 + ε2(1 + ε)) and \widetildeO(T^\f rac1+ 2ε1 + 3 ε) respectively. In particular, they guarantee sublinear regret as long as the noise has finite variance. We also present empirical results showin g that our algorithms achieve a better performance than the current state of the art for bounded noise when the L_∞bound on the noise is large yet the 1 + εmome nt of the noise is small.
****
Extended and Unscented Kitchen Sinks
Edwin Bonilla, Daniel Steinberg, Alistair Reid
We propose a scalable multiple-output generalization of unscented and extended G aussian processes. These algorithms have been designed to handle general likelih ood models by linearizing them using a Taylor series or the Unscented Transform in a variational inference framework. We build upon random feature approximation s of Gaussian process covariance functions and show that, on small-scale single- task problems, our methods can attain similar performance as the original algori thms while having less computational cost. We also evaluate our methods at a lar ger scale on MNIST and on a seismic inversion which is inherently a multi-task p roblem.
****
Matrix Eigen-decomposition via Doubly Stochastic Riemannian Optimization
Zhiqiang Xu, Peilin Zhao, Jianneng Cao, Xiaoli Li
Matrix eigen-decomposition is a classic and long-standing problem that plays a f undamental role in scientific computing and machine learning. Despite some exist ing algorithms for this inherently non-convex problem, the study remains inadequ ate for the need of large data nowadays. To address this gap, we propose a Doubl

y Stochastic Riemannian Gradient EIGenSolver, DSRG-EIGS, where the double stocha sticity comes from the generalization of the stochastic Euclidean gradient ascen t and the stochastic Euclidean coordinate ascent to Riemannian manifolds. As a r esult, it induces a greatly reduced complexity per iteration, enables the algori thm to completely avoid the matrix inversion, and consequently makes it well-sui ted to large-scale applications. We theoretically analyze its convergence proper ties and empirically validate it on real-world datasets. Encouraging experimenta l results demonstrate its advantages over the deterministic counterparts.
****
Recommendations as Treatments: Debiasing Learning and Evaluation
Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, Thorsten Joac hims
Most data for evaluating and training recommender systems is subject to selectio n biases, either through self-selection by the users or through the actions of t he recommendation system itself. In this paper, we provide a principled approach to handle selection biases by adapting models and estimation techniques from ca usal inference. The approach leads to unbiased performance estimators despite bi ased data, and to a matrix factorization method that provides substantially impr oved prediction performance on real-world data. We theoretically and empirically characterize the robustness of the approach, and find that it is highly practic al and scalable.
****
ForecastICU: A Prognostic Decision Support System for Timely Prediction of Inten sive Care Unit Admission
Jinsung Yoon, Ahmed Alaa, Scott Hu, Mihaela Schaar
We develop ForecastICU: a prognostic decision support system that monitors hospi talized patients and prompts alarms for intensive care unit (ICU) admissions. Fo recastICU is first trained in an offline stage by constructing a Bayesian belief system that corresponds to its belief about how trajectories of physiological d ata streams of the patient map to a clinical status. After that, ForecastICU mon itors a new patient in real-time by observing her physiological data stream, upd ating its belief about her status over time, and prompting an alarm whenever its belief process hits a predefined threshold (confidence). Using a real-world dat aset obtained from UCLA Ronald Reagan Medical Center, we show that ForecastICU c an predict ICU admissions 9 hours before a physician’s decision (for a sensitivi ty of 40% and a precision of 50%). Also, ForecastICU performs consistently bette r than other state-of-the-art machine learning algorithms in terms of sensitivit y, precision, and timeliness: it can predict ICU admissions 3 hours earlier, and offers a 7.8% gain in sensitivity and a 5.1% gain in precision compared to the best state-of-the-art algorithm. Moreover, ForecastICU offers an area under curv e (AUC) gain of 22.3% compared to the Rothman index, which is the currently depl oyed technology in most hospital wards.
****
An optimal algorithm for the Thresholding Bandit Problem
Andrea Locatelli, Maurilio Gutzeit, Alexandra Carpentier
We study a specific combinatorial pure exploration stochastic bandit problem whe re the learner aims at finding the set of arms whose means are above a given thr eshold, up to a given precision, and for a fixed time horizon. We propose a para meter-free algorithm based on an original heuristic, and prove that it is optima l for this problem by deriving matching upper and lower bounds. To the best of o ur knowledge, this is the first non-trivial pure exploration setting with fixed budget for which provably optimal strategies are constructed.
****
Fast Parameter Inference in Nonlinear Dynamical Systems using Iterative Gradient Matching
Mu Niu, Simon Rogers, Maurizio Filippone, Dirk Husmeier
Parameter inference in mechanistic models of coupled differential equations is a topical and challenging problem. We propose a new method based on kernel ridge regression and gradient matching, and an objective function that simultaneously encourages goodness of fit and penalises inconsistencies with the differential e

quations. Fast minimisation is achieved by exploiting partial convexity inherent in this function, and setting up an iterative algorithm in the vein of the EM algorithm. An evaluation of the proposed method on various benchmark data suggests that it compares favourably with state-of-the-art alternatives.

****

Structured and Efficient Variational Deep Learning with Matrix Gaussian Posteriors

Christos Louizos, Max Welling

We introduce a variational Bayesian neural network where the parameters are governed via a probability distribution on random matrices. Specifically, we employ a matrix variate Gaussian (Gupta & Nagar ’99) parameter posterior distribution where we explicitly model the covariance among the input and output dimensions of each layer. Furthermore, with approximate covariance matrices we can achieve a more efficient way to represent those correlations that is also cheaper than fully factorized parameter posteriors. We further show that with the “local reprarametrization trick" (Kingma & Welling ’15) on this posterior distribution we arrive at a Gaussian Process (Rasmussen ’06) interpretation of the hidden units in each layer and we, similarly with (Gal & Ghahramani ’15), provide connections with deep Gaussian processes. We continue in taking advantage of this duality and incorporate “pseudo-data” (Snelson & Ghahramani ’05) in our model, which in turn allows for more efficient posterior sampling while maintaining the properties of the original model. The validity of the proposed approach is verified through extensive experiments.

****

Learning Granger Causality for Hawkes Processes

Hongteng Xu, Mehrdad Farajtabar, Hongyuan Zha

Learning Granger causality for general point processes is a very challenging task. We propose an effective method learning Granger causality for a special but significant type of point processes — Hawkes processes. Focusing on Hawkes processes, we reveal the relationship between Hawkes process’s impact functions and its Granger causality graph. Specifically, our model represents impact functions using a series of basis functions and recovers the Granger causality graph via group sparsity of the impact functions’ coefficients. We propose an effective learning algorithm combining a maximum likelihood estimator (MLE) with a sparse-group-lasso (SGL) regularizer. Additionally, the pairwise similarity between the dimensions of the process is considered when their clustering structure is available. We analyze our learning method and discuss the selection of the basis functions. Experiments on synthetic data and real-world data show that our method can learn the Granger causality graph and the triggering patterns of Hawkes processes simultaneously.

****

Neural Variational Inference for Text Processing

Yishu Miao, Lei Yu, Phil Blunsom

Recent advances in neural variational inference have spawned a renaissance in deep latent variable models. In this paper we introduce a generic variational inference framework for generative and conditional models of text. While traditional variational methods derive an analytic approximation for the intractable distributions over latent variables, here we construct an inference network conditioned on the discrete text input to provide the variational distribution. We validate this framework on two very different text modelling applications, generative document modelling and supervised question answering. Our neural variational document model combines a continuous stochastic document representation with a bag-of-words generative model and achieves the lowest reported perplexities on two standard test corpora. The neural answer selection model employs a stochastic representation layer within an attention mechanism to extract the semantics between a question and answer pair. On two question answering benchmarks this model exceeds all previous published benchmarks.

****

Dictionary Learning for Massive Matrix Factorization

Arthur Mensch, Julien Mairal, Bertrand Thirion, Gael Varoquaux

Sparse matrix factorization is a popular tool to obtain interpretable data decom
positions, which are also effective to perform data completion or denoising. Its
 applicability to large datasets has been addressed with online and randomized m
ethods, that reduce the complexity in one of the matrix dimension, but not in bo
th of them. In this paper, we tackle very large matrices in both dimensions. We
propose a new factorization method that scales gracefully to terabyte-scale data
sets. Those could not be processed by previous algorithms in a reasonable amount
 of time. We demonstrate the efficiency of our approach on massive functional Ma
gnetic Resonance Imaging (fMRI) data, and on matrix completion problems for reco
mmender systems, where we obtain significant speed-ups compared to state-of-the
art coordinate descent methods.
****

Pixel Recurrent Neural Networks
Aäron van den Oord, Nal Kalchbrenner, Koray Kavukcuoglu
Modeling the distribution of natural images is a landmark problem in unsupervise
d learning. This task requires an image model that is at once expressive, tracta
ble and scalable. We present a deep neural network that sequentially predicts th
e pixels in an image along the two spatial dimensions. Our method models the dis
crete probability of the raw pixel values and encodes the complete set of depend
encies in the image. Architectural novelties include fast two-dimensional recurr
ent layers and an effective use of residual connections in deep recurrent networ
ks. We achieve log-likelihood scores on natural images that are considerably bet
ter than the previous state of the art. Our main results also provide benchmarks
 on the diverse ImageNet dataset. Samples generated from the model appear crisp,
 varied and globally coherent.
****

Why Most Decisions Are Easy in Tetris—And Perhaps in Other Sequential Decision P
roblems, As Well
Özgür ■im■ek, Simón Algorta, Amit Kothiyal
We examined the sequence of decision problems that are encountered in the game o
f Tetris and found that most of the problems are easy in the following sense: On
e can choose well among the available actions without knowing an evaluation func
tion that scores well in the game. This is a consequence of three conditions tha
t are prevalent in the game: simple dominance, cumulative dominance, and noncomp
ensation. These conditions can be exploited to develop faster and more effective
 learning algorithms. In addition, they allow certain types of domain knowledge
to be incorporated with ease into a learning algorithm. Among the sequential dec
ision problems we encounter, it is unlikely that Tetris is unique or rare in hav
ing these properties.
****

Gaussian quadrature for matrix inverse forms with applications
Chengtao Li, Suvrit Sra, Stefanie Jegelka
We present a framework for accelerating a spectrum of machine learning algorithm
s that require computation of \emphbilinear inverse forms u^T A^-1u, where A is
a positive definite matrix and u a given vector. Our framework is built on Gauss
-type quadrature and easily scales to large, sparse matrices. Further, it allows
 retrospective computation of lower and upper bounds on u^T A^-1u, which in turn
 accelerates several algorithms. We prove that these bounds tighten iteratively
and converge at a linear (geometric) rate. To our knowledge, ours is the first w
ork to demonstrate these key properties of Gauss-type quadrature, which is a cla
ssical and deeply studied topic. We illustrate empirical consequences of our res
ults by using quadrature to accelerate machine learning tasks involving determin
antal point processes and submodular optimization, and observe tremendous speedu
ps in several instances.
****

Train and Test Tightness of LP Relaxations in Structured Prediction
Ofer Meshi, Mehrdad Mahdavi, Adrian Weller, David Sontag
Structured prediction is used in areas such as computer vision and natural langu
age processing to predict structured outputs such as segmentations or parse tree
s. In these settings, prediction is performed by MAP inference or, equivalently,

by solving an integer linear program. Because of the complex scoring functions required to obtain accurate predictions, both learning and inference typically require the use of approximate solvers. We propose a theoretical explanation to the striking observation that approximations based on linear programming (LP) relaxations are often tight on real-world instances. In particular, we show that learning with LP relaxed inference encourages integrality of training instances, and that tightness generalizes from train to test data.
****

Stochastic Optimization for Multiview Representation Learning using Partial Least Squares
Raman Arora, Poorya Mianjy, Teodor Marinov
Partial Least Squares (PLS) is a ubiquitous statistical technique for bilinear factor analysis. It is used in many data analysis, machine learning, and information retrieval applications to model the covariance structure between a pair of data matrices. In this paper, we consider PLS for representation learning in a multiview setting where we have more than one view in data at training time. Furthermore, instead of framing PLS as a problem about a fixed given data set, we argue that PLS should be studied as a stochastic optimization problem, especially in a "big data" setting, with the goal of optimizing a population objective based on sample. This view suggests using Stochastic Approximation (SA) approaches, such as Stochastic Gradient Descent (SGD) and enables a rigorous analysis of their benefits. In this paper, we develop SA approaches to PLS and provide iteration complexity bounds for the proposed algorithms.
****

Hierarchical Compound Poisson Factorization
Mehmet Basbug, Barbara Engelhardt
Non-negative matrix factorization models based on a hierarchical Gamma-Poisson structure capture user and item behavior effectively in extremely sparse data sets, making them the ideal choice for collaborative filtering applications. Hierarchical Poisson factorization (HPF) in particular has proved successful for scalable recommendation systems with extreme sparsity. HPF, however, suffers from a tight coupling of sparsity model (absence of a rating) and response model (the value of the rating), which limits the expressiveness of the latter. Here, we introduce hierarchical compound Poisson factorization (HCPF) that has the favorable Gamma-Poisson structure and scalability of HPF to high-dimensional extremely sparse matrices. More importantly, HCPF decouples the sparsity model from the response model, allowing us to choose the most suitable distribution for the response. HCPF can capture binary, non-negative discrete, non-negative continuous, and zero-inflated continuous responses. We compare HCPF with HPF on nine discrete and three continuous data sets and conclude that HCPF captures the relationship between sparsity and response better than HPF.
****

Opponent Modeling in Deep Reinforcement Learning
He He, Jordan Boyd-Graber, Kevin Kwok, Hal Daumé III
Opponent modeling is necessary in multi-agent settings where secondary agents with competing goals also adapt their strategies, yet it remains challenging because of strategies’ complex interaction and the non-stationary nature. Most previous work focuses on developing probabilistic models or parameterized strategies for specific applications. Inspired by the recent success of deep reinforcement learning, we present neural-based models that jointly learn a policy and the behavior of opponents. Instead of explicitly predicting the opponent’s action, we encode observation of the opponents into a deep Q-Network (DQN), while retaining explicit modeling under multitasking. By using a Mixture-of-Experts architecture, our model automatically discovers different strategy patterns of opponents even without extra supervision. We evaluate our models on a simulated soccer game and a popular trivia game, showing superior performance over DQN and its variants.
****

No penalty no tears: Least squares in high-dimensional linear models
Xiangyu Wang, David Dunson, Chenlei Leng
Ordinary least squares (OLS) is the default method for fitting linear models, bu

t is not applicable for problems with dimensionality larger than the sample size . For these problems, we advocate the use of a generalized version of OLS motiva ted by ridge regression, and propose two novel three-step algorithms involving l east squares fitting and hard thresholding. The algorithms are methodologically simple to understand intuitively, computationally easy to implement efficiently, and theoretically appealing for choosing models consistently. Numerical exercis es comparing our methods with penalization-based approaches in simulations and d ata analyses illustrate the great potential of the proposed algorithms.
****
SDNA: Stochastic Dual Newton Ascent for Empirical Risk Minimization
Zheng Qu, Peter Richtarik, Martin Takac, Olivier Fercoq
We propose a new algorithm for minimizing regularized empirical loss: Stochastic Dual Newton Ascent (SDNA). Our method is dual in nature: in each iteration we u pdate a random subset of the dual variables. However, unlike existing methods su ch as stochastic dual coordinate ascent, SDNA is capable of utilizing all local curvature information contained in the examples, which leads to striking improve ments in both theory and practice – sometimes by orders of magnitude. In the spe cial case when an L2-regularizer is used in the primal, the dual problem is a co ncave quadratic maximization problem plus a separable term. In this regime, SDNA in each step solves a proximal subproblem involving a random principal submatri x of the Hessian of the quadratic function; whence the name of the method.
****
On Graduated Optimization for Stochastic Non-Convex Problems
Elad Hazan, Kfir Yehuda Levy, Shai Shalev-Shwartz
The graduated optimization approach, also known as the continuation method, is a popular heuristic to solving non-convex problems that has received renewed inte rest over the last decade.Despite being popular, very little is known in terms o f its theoretical convergence analysis. In this paper we describe a new first-or der algorithm based on graduated optimization and analyze its performance. We ch aracterize a family of non-convex functions for which this algorithm provably co nverges to a global optimum. In particular, we prove that the algorithm converge s to an ε-approximate solution within $O(1 / \varepsilon^2)$ gradient-based steps. We extend our algorithm and analysis to the setting of stochastic non-convex optimization with noisy gradient feedback, attaining the same convergence rate. Additionally , we discuss the setting of “zero-order optimization", and devise a variant of o ur algorithm which converges at rate of $O(d^2/ \varepsilon^4)$.
****
Meta-Learning with Memory-Augmented Neural Networks
Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, Timothy Lillicr ap
Despite recent breakthroughs in the applications of deep neural networks, one se tting that presents a persistent challenge is that of "one-shot learning." Tradi tional gradient-based networks require a lot of data to learn, often through ext ensive iterative training. When new data is encountered, the models must ineffic iently relearn their parameters to adequately incorporate the new information wi thout catastrophic interference. Architectures with augmented memory capacities, such as Neural Turing Machines (NTMs), offer the ability to quickly encode and retrieve new information, and hence can potentially obviate the downsides of con ventional models. Here, we demonstrate the ability of a memory-augmented neural network to rapidly assimilate new data, and leverage this data to make accurate predictions after only a few samples. We also introduce a new method for accessi ng an external memory that focuses on memory content, unlike previous methods th at additionally use memory location-based focusing mechanisms.
****
The knockoff filter for FDR control in group-sparse and multitask regression
Ran Dai, Rina Barber
We propose the group knockoff filter, a method for false discovery rate control in a linear regression setting where the features are grouped, and we would like to select a set of relevant groups which have a nonzero effect on the response. By considering the set of true and false discoveries at the group level, this m

ethod gains power relative to sparse regression methods. We also apply our metho d to the multitask regression problem where multiple response variables share si milar sparsity patterns across the set of possible features. Empirically, the gr oup knockoff filter successfully controls false discoveries at the group level i n both settings, with substantially more discoveries made by leveraging the grou p structure.
****

Softened Approximate Policy Iteration for Markov Games
Julien Pérolat, Bilal Piot, Matthieu Geist, Bruno Scherrer, Olivier Pietquin
This paper reports theoretical and empirical investigations on the use of quasi-Newton methods to minimize the Optimal Bellman Residual (OBR) of zero-sum two-pl ayer Markov Games. First, it reveals that state-of-the-art algorithms can be der ived by the direct application of Newton’s method to different norms of the OBR. More precisely, when applied to the norm of the OBR, Newton’s method results in the Bellman Residual Minimization Policy Iteration (BRMPI) and, when applied to the norm of the Projected OBR (POBR), it results into the standard Least Square s Policy Iteration (LSPI) algorithm. Consequently, new algorithms are proposed, making use of quasi-Newton methods to minimize the OBR and the POBR so as to tak e benefit of enhanced empirical performances at low cost. Indeed, using a quasi-Newton method approach introduces slight modifications in term of coding of LSPI and BRMPI but improves significantly both the stability and the performance of those algorithms. These phenomena are illustrated on an experiment conducted on artificially constructed games called Garnets.
****

Stochastic Block BFGS: Squeezing More Curvature out of Data
Robert Gower, Donald Goldfarb, Peter Richtarik
We propose a novel limited-memory stochastic block BFGS update for incorporating enriched curvature information in stochastic approximation methods. In our meth od, the estimate of the inverse Hessian matrix that is maintained by it, is upda ted at each iteration using a sketch of the Hessian, i.e., a randomly generated compressed form of the Hessian. We propose several sketching strategies, present a new quasi-Newton method that uses stochastic block BFGS updates combined with the variance reduction approach SVRG to compute batch stochastic gradients, and prove linear convergence of the resulting method. Numerical tests on large-scal e logistic regression problems reveal that our method is more robust and substan tially outperforms current state-of-the-art methods.
****

Differential Geometric Regularization for Supervised Learning of Classifiers
Qinxun Bai, Steven Rosenberg, Zheng Wu, Stan Sclaroff
We study the problem of supervised learning for both binary and multiclass class ification from a unified geometric perspective. In particular, we propose a geom etric regularization technique to find the submanifold corresponding to an estim ator of the class probability P(y|\vec x). The regularization term measures the volume of this submanifold, based on the intuition that overfitting produces rap id local oscillations and hence large volume of the estimator. This technique ca n be applied to regularize any classification function that satisfies two requir ements: firstly, an estimator of the class probability can be obtained; secondly , first and second derivatives of the class probability estimator can be calcula ted. In experiments, we apply our regularization technique to standard loss func tions for classification, our RBF-based implementation compares favorably to wid ely used regularization methods for both binary and multiclass classification.
****

Exploiting Cyclic Symmetry in Convolutional Neural Networks
Sander Dieleman, Jeffrey De Fauw, Koray Kavukcuoglu
Many classes of images exhibit rotational symmetry. Convolutional neural network s are sometimes trained using data augmentation to exploit this, but they are st ill required to learn the rotation equivariance properties from the data. Encodi ng these properties into the network architecture, as we are already used to doi ng for translation equivariance by using convolutional layers, could result in a more efficient use of the parameter budget by relieving the model from learning

them. We introduce four operations which can be inserted into neural network models as layers, and which can be combined to make these models partially equivariant to rotations. They also enable parameter sharing across different orientations. We evaluate the effect of these architectural modifications on three datasets which exhibit rotational symmetry and demonstrate improved performance with smaller models.

****

Graying the black box: Understanding DQNs

Tom Zahavy, Nir Ben-Zrihem, Shie Mannor

In recent years there is a growing interest in using deep representations for reinforcement learning. In this paper, we present a methodology and tools to analyze Deep Q-networks (DQNs) in a non-blind matter. Using our tools we reveal that the features learned by DQNs aggregate the state space in a hierarchical fashion, explaining its success. Moreover we are able to understand and describe the policies learned by DQNs for three different Atari2600 games and suggest ways to interpret, debug and optimize of deep neural networks in Reinforcement Learning.

****

The Sum-Product Theorem: A Foundation for Learning Tractable Models

Abram Friesen, Pedro Domingos

Inference in expressive probabilistic models is generally intractable, which makes them difficult to learn and limits their applicability. Sum-product networks are a class of deep models where, surprisingly, inference remains tractable even when an arbitrary number of hidden layers are present. In this paper, we generalize this result to a much broader set of learning problems: all those where inference consists of summing a function over a semiring. This includes satisfiability, constraint satisfaction, optimization, integration, and others. In any semiring, for summation to be tractable it suffices that the factors of every product have disjoint scopes. This unifies and extends many previous results in the literature. Enforcing this condition at learning time thus ensures that the learned models are tractable. We illustrate the power and generality of this approach by applying it to a new type of structured prediction problem: learning a nonconvex function that can be globally optimized in polynomial time. We show empirically that this greatly outperforms the standard approach of learning without regard to the cost of optimization.

****

Pareto Frontier Learning with Expensive Correlated Objectives

Amar Shah, Zoubin Ghahramani

There has been a surge of research interest in developing tools and analysis for Bayesian optimization, the task of finding the global maximizer of an unknown, expensive function through sequential evaluation using Bayesian decision theory. However, many interesting problems involve optimizing multiple, expensive to evaluate objectives simultaneously, and relatively little research has addressed this setting from a Bayesian theoretic standpoint. A prevailing choice when tackling this problem, is to model the multiple objectives as being independent, typically for ease of computation. In practice, objectives are correlated to some extent. In this work, we incorporate the modelling of inter-task correlations, developing an approximation to overcome intractable integrals. We illustrate the power of modelling dependencies between objectives on a range of synthetic and real world multi-objective optimization problems.

****

Asynchronous Methods for Deep Reinforcement Learning

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, Koray Kavukcuoglu

We propose a conceptually simple and lightweight framework for deep reinforcement learning that uses asynchronous gradient descent for optimization of deep neural network controllers. We present asynchronous variants of four standard reinforcement learning algorithms and show that parallel actor-learners have a stabilizing effect on training allowing all four methods to successfully train neural network controllers. The best performing method, an asynchronous variant of actor-critic, surpasses the current state-of-the-art on the Atari domain while traini

ng for half the time on a single multi-core CPU instead of a GPU. Furthermore, w e show that asynchronous actor-critic succeeds on a wide variety of continuous m otor control problems as well as on a new task of navigating random 3D mazes usi ng a visual input.

****

A Simple and Strongly-Local Flow-Based Method for Cut Improvement

Nate Veldt, David Gleich, Michael Mahoney

Many graph-based learning problems can be cast as finding a good set of vertices nearby a seed set, and a powerful methodology for these problems is based on mi nimum cuts and maximum flows. We introduce and analyze a new method for locally-biased graph-based learning called SimpleLocal, which finds good conductance cut s near a set of seed vertices. An important feature of our algorithm is that it is strongly-local, meaning it does not need to explore the entire graph to find cuts that are locally optimal. This method is related to other strongly-local fl ow-based methods, but it enables a simple implementation. We also show how it ac hieves localization through an implicit l1-norm penalty term. As a flow-based me thod, our algorithm exhibits several advantages in terms of cut optimality and a ccurate identification of target regions in a graph. We demonstrate the power of SimpleLocal solving segmentation problems on a 467 million edge graph based on an MRI scan.

****

Nonlinear Statistical Learning with Truncated Gaussian Graphical Models

Qinliang Su, Xuejun Liao, Changyou Chen, Lawrence Carin

We introduce the truncated Gaussian graphical model (TGGM) as a novel framework for designing statistical models for nonlinear learning. A TGGM is a Gaussian gr aphical model (GGM) with a subset of variables truncated to be nonnegative. The truncated variables are assumed latent and integrated out to induce a marginal m odel. We show that the variables in the marginal model are non-Gaussian distribu ted and their expected relations are nonlinear. We use expectation-maximization to break the inference of the nonlinear model into a sequence of TGGM inference problems, each of which is efficiently solved by using the properties and numeri cal methods of multivariate Gaussian distributions. We use the TGGM to design mo dels for nonlinear regression and classification, with the performances of these models demonstrated on extensive benchmark datasets and compared to state-of-th e-art competing results.

****

Barron and Cover’s Theory in Supervised Learning and its Application to Lasso

Masanori Kawakita, Jun’ichi Takeuchi

We study Barron and Cover’s theory (BC theory) in supervised learning. The origi nal BC theory can be applied to supervised learning only approximately and limit edly. Though Barron (2008) and Chatterjee and Barron (2014) succeeded in removin g the approximation, their idea cannot be essentially applied to supervised lear ning in general. By solving this issue, we propose an extension of BC theory to supervised learning. The extended theory has several advantages inherited from t he original BC theory. First, it holds for finite sample number n. Second, it re quires remarkably few assumptions. Third, it gives a justification of the MDL pr inciple in supervised learning. We also derive new risk and regret bounds of las so with random design as its application. The derived risk bound hold for any fi nite n without boundedness of features in contrast to past work. Behavior of the regret bound is investigated by numerical simulations. We believe that this is the first extension of BC theory to general supervised learning without approxim ation.

****

Nonparametric Canonical Correlation Analysis

Tomer Michaeli, Weiran Wang, Karen Livescu

Canonical correlation analysis (CCA) is a classical representation learning tech nique for finding correlated variables in multi-view data. Several nonlinear ext ensions of the original linear CCA have been proposed, including kernel and deep neural network methods. These approaches seek maximally correlated projections among families of functions, which the user specifies (by choosing a kernel or n

eural network structure), and are computationally demanding. Interestingly, the theory of nonlinear CCA, without functional restrictions, had been studied in th e population setting by Lancaster already in the 1950s, but these results have n ot inspired practical algorithms. We revisit Lancaster’s theory to devise a prac tical algorithm for nonparametric CCA (NCCA). Specifically, we show that the sol ution can be expressed in terms of the singular value decomposition of a certain operator associated with the joint density of the views. Thus, by estimating th e population density from data, NCCA reduces to solving an eigenvalue system, su perficially like kernel CCA but, importantly, without requiring the inversion of any kernel matrix. We also derive a partially linear CCA (PLCCA) variant in whi ch one of the views undergoes a linear projection while the other is nonparametr ic. Using a kernel density estimate based on a small number of nearest neighbors , our NCCA and PLCCA algorithms are memory-efficient, often run much faster, and perform better than kernel CCA and comparable to deep CCA.

****

BISTRO: An Efficient Relaxation-Based Method for Contextual Bandits

Alexander Rakhlin, Karthik Sridharan

We present efficient algorithms for the problem of contextual bandits with i.i.d . covariates, an arbitrary sequence of rewards, and an arbitrary class of polici es. Our algorithm BISTRO requires d calls to the empirical risk minimization (ER M) oracle per round, where d is the number of actions. The method uses unlabeled data to make the problem computationally simple. When the ERM problem itself is computationally hard, we extend the approach by employing multiplicative approx imation algorithms for the ERM. The integrality gap of the relaxation only enter s in the regret bound rather than the benchmark. Finally, we show that the adver sarial version of the contextual bandit problem is learnable (and efficient) whe never the full-information supervised online learning problem has a non-trivial regret bound (and efficient).

****

Associative Long Short-Term Memory

Ivo Danihelka, Greg Wayne, Benigno Uria, Nal Kalchbrenner, Alex Graves

We investigate a new method to augment recurrent neural networks with extra memo ry without increasing the number of network parameters. The system has an associ ative memory based on complex-valued vectors and is closely related to Holograph ic Reduced Representations and Long Short-Term Memory networks. Holographic Redu ced Representations have limited capacity: as they store more information, each retrieval becomes noisier due to interference. Our system in contrast creates re dundant copies of stored information, which enables retrieval with reduced noise . Experiments demonstrate faster learning on multiple memorization tasks.

****

Dueling Network Architectures for Deep Reinforcement Learning

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, Nando Freitas

In recent years there have been many successes of using deep representations in reinforcement learning. Still, many of these applications use conventional archi tectures, such as convolutional networks, LSTMs, or auto-encoders. In this paper , we present a new neural network architecture for model-free reinforcement lear ning. Our dueling network represents two separate estimators: one for the state value function and one for the state-dependent action advantage function. The ma in benefit of this factoring is to generalize learning across actions without im posing any change to the underlying reinforcement learning algorithm. Our result s show that this architecture leads to better policy evaluation in the presence of many similar-valued actions. Moreover, the dueling architecture enables our R L agent to outperform the state-of-the-art on the Atari 2600 domain.

****

Persistence weighted Gaussian kernel for topological data analysis

Genki Kusano, Yasuaki Hiraoka, Kenji Fukumizu

Topological data analysis (TDA) is an emerging mathematical concept for characte rizing shapes in complex data. In TDA, persistence diagrams are widely recognize d as a useful descriptor of data, and can distinguish robust and noisy topologic al properties. This paper proposes a kernel method on persistence diagrams to de

velop a statistical framework in TDA. The proposed kernel satisfies the stabilit y property and provides explicit control on the effect of persistence. Furthermo re, the method allows a fast approximation technique. The method is applied into practical data on proteins and oxide glasses, and the results show the advantag e of our method compared to other relevant methods on persistence diagrams.
****

Learning Convolutional Neural Networks for Graphs
Mathias Niepert, Mohamed Ahmed, Konstantin Kutzkov
Numerous important problems can be framed as learning from graph data. We propos e a framework for learning convolutional neural networks for arbitrary graphs. T hese graphs may be undirected, directed, and with both discrete and continuous n ode and edge attributes. Analogous to image-based convolutional networks that op erate on locally connected regions of the input, we present a general approach t o extracting locally connected regions from graphs. Using established benchmark data sets, we demonstrate that the learned feature representations are competiti ve with state of the art graph kernels and that their computation is highly effi cient.
****

Persistent RNNs: Stashing Recurrent Weights On-Chip
Greg Diamos, Shubho Sengupta, Bryan Catanzaro, Mike Chrzanowski, Adam Coates, Er ich Elsen, Jesse Engel, Awni Hannun, Sanjeev Satheesh
This paper introduces a new technique for mapping Deep Recurrent Neural Networks (RNN) efficiently onto GPUs. We show how it is possi- ble to achieve substantia lly higher computational throughput at low mini-batch sizes than direct implemen tations of RNNs based on matrix multiplications. The key to our approach is the use of persistent computational kernels that exploit the GPU’s inverted memory h ierarchy to reuse network weights over multiple timesteps. Our initial implement ation sustains 2.8 TFLOP/s at a mini-batch size of 4 on an NVIDIA TitanX GPU. Th is provides a 16x reduction in activation memory footprint, enables model traini ng with 12x more parameters on the same hardware, allows us to strongly scale RN N training to 128 GPUs, and allows us to efficiently explore end-to-end speech r ecognition models with over 100 layers.
****

Recurrent Orthogonal Networks and Long-Memory Tasks
Mikael Henaff, Arthur Szlam, Yann LeCun
Although RNNs have been shown to be power- ful tools for processing sequential d ata, finding architectures or optimization strategies that al- low them to model very long term dependencies is still an active area of research. In this work, we carefully analyze two synthetic datasets orig- inally outlined in (Hochreiter & Schmidhuber, 1997) which are used to evaluate the ability of RNNs to store in formation over many time steps. We explicitly construct RNN solutions to these p roblems, and using these constructions, illumi- nate both the problems themselve s and the way in which RNNs store different types of information in their hidden states. These constructions fur- thermore explain the success of recent methods that specify unitary initializations or constraints on the transition matrices.
****

The Arrow of Time in Multivariate Time Series
Stefan Bauer, Bernhard Schölkopf, Jonas Peters
We prove that a time series satisfying a (linear) multivariate autoregressive mo ving average (VARMA) model satisfies the same model assumption in the reversed t ime direction, too, if all innovations are normally distributed. This reversibil ity breaks down if the innovations are non-Gaussian. This means that under the a ssumption of a VARMA process with non-Gaussian noise, the arrow of time becomes detectable. Our work thereby provides a theoretic justification of an algorithm that has been used for inferring the direction of video snippets. We present a s lightly modified practical algorithm that estimates the time direction for a giv en sample and prove its consistency. We further investigate how the performance of the algorithm depends on sample size, number of dimensions of the time series and the order of the process. An application to real world data from economics shows that considering multivariate processes instead of univariate processes ca

n be beneficial for estimating the time direction. Our result extends earlier wo rk on univariate time series. It relates to the concept of causal inference, whe re recent methods exploit non-Gaussianity of the error terms for causal structur e learning.
****
Mixture Proportion Estimation via Kernel Embeddings of Distributions
Harish Ramaswamy, Clayton Scott, Ambuj Tewari
Mixture proportion estimation (MPE) is the problem of estimating the weight of a component distribution in a mixture, given samples from the mixture and compone nt. This problem constitutes a key part in many "weakly supervised learning" pro blems like learning with positive and unlabelled samples, learning with label no ise, anomaly detection and crowdsourcing. While there have been several methods proposed to solve this problem, to the best of our knowledge no efficient algori thm with a proven convergence rate towards the true proportion exists for this p roblem. We fill this gap by constructing a provably correct algorithm for MPE, a nd derive convergence rates under certain assumptions on the distribution. Our m ethod is based on embedding distributions onto an RKHS, and implementing it only requires solving a simple convex quadratic programming problem a few times. We run our algorithm on several standard classification datasets, and demonstrate t hat it performs comparably to or better than other algorithms on most datasets.
****
Fast DPP Sampling for Nystrom with Application to Kernel Methods
Chengtao Li, Stefanie Jegelka, Suvrit Sra
The Nystrom method has long been popular for scaling up kernel methods. Its theo retical guarantees and empirical performance rely critically on the quality of t he landmarks selected. We study landmark selection for Nystrom using Determinant al Point Processes (DPPs), discrete probability models that allow tractable gene ration of diverse samples. We prove that landmarks selected via DPPs guarantee b ounds on approximation errors; subsequently, we analyze implications for kernel ridge regression. Contrary to prior reservations due to cubic complexity of DPP sampling, we show that (under certain conditions) Markov chain DPP sampling requ ires only linear time in the size of the data. We present several empirical resu lts that support our theoretical analysis, and demonstrate the superior performa nce of DPP-based landmark selection compared with existing approaches.
****
Complex Embeddings for Simple Link Prediction
Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, Guillaume Bouch ard
In statistical relational learning, the link prediction problem is key to automa tically understand the structure of large knowledge bases. As in previous studie s, we propose to solve this problem through latent factorization. However, here we make use of complex valued embeddings. The composition of complex embeddings can handle a large variety of binary relations, among them symmetric and antisym metric relations. Compared to state-of-the-art models such as Neural Tensor Netw ork and Holographic Embeddings, our approach based on complex embeddings is argu ably simpler, as it only uses the Hermitian dot product, the complex counterpart of the standard dot product between real vectors. Our approach is scalable to l arge datasets as it remains linear in both space and time, while consistently ou tperforming alternative approaches on standard link prediction benchmarks.
****
Interactive Bayesian Hierarchical Clustering
Sharad Vikram, Sanjoy Dasgupta
Clustering is a powerful tool in data analysis, but it is often difficult to fin d a grouping that aligns with a user’s needs. To address this, several methods i ncorporate constraints obtained from users into clustering algorithms, but unfor tunately do not apply to hierarchical clustering. We design an interactive Bayes ian algorithm that incorporates user interaction into hierarchical clustering wh ile still utilizing the geometry of the data by sampling a constrained posterior distribution over hierarchies. We also suggest several ways to intelligently qu ery a user. The algorithm, along with the querying schemes, shows promising resu

lts on real data.
****
A Convolutional Attention Network for Extreme Summarization of Source Code
Miltiadis Allamanis, Hao Peng, Charles Sutton
Attention mechanisms in neural networks have proved useful for problems in which the input and output do not have fixed dimension. Often there exist features that are locally translation invariant and would be valuable for directing the model's attention, but previous attentional architectures are not constructed to learn such features specifically. We introduce an attentional neural network that employs convolution on the input tokens to detect local time-invariant and long-range topical attention features in a context-dependent way. We apply this architecture to the problem of extreme summarization of source code snippets into short, descriptive function name-like summaries. Using those features, the model sequentially generates a summary by marginalizing over two attention mechanisms: one that predicts the next summary token based on the attention weights of the input tokens and another that is able to copy a code token as-is directly into the summary. We demonstrate our convolutional attention neural network's performance on 10 popular Java projects showing that it achieves better performance compared to previous attentional mechanisms.
****
How to Fake Multiply by a Gaussian Matrix
Michael Kapralov, Vamsi Potluru, David Woodruff
Have you ever wanted to multiply an n \times d matrix X, with n ■d, on the left by an m \times n matrix \tilde G of i.i.d. Gaussian random variables, but could not afford to do it because it was too slow? In this work we propose a new randomized m \times n matrix T, for which one can compute T ·X in only O(nnz(X)) + \tilde O(m^1.5 ·d^3) time, for which the total variation distance between the distributions T ·X and \tilde G ·X is as small as desired, i.e., less than any positive constant. Here nnz(X) denotes the number of non-zero entries of X. Assuming nnz(X) ■m^1.5 ·d^3, this is a significant savings over the naïve O(nnz(X) m) time to compute \tilde G ·X. Moreover, since the total variation distance is small, we can provably use T ·X in place of \tilde G ·X in any application and have the same guarantees as if we were using \tilde G ·X, up to a small positive constant in error probability. We apply this transform to nonnegative matrix factorization (NMF) and support vector machines (SVM).
****
Differentially Private Chi-Squared Hypothesis Testing: Goodness of Fit and Independence Testing
Marco Gaboardi, Hyun Lim, Ryan Rogers, Salil Vadhan
Hypothesis testing is a useful statistical tool in determining whether a given model should be rejected based on a sample from the population. Sample data may contain sensitive information about individuals, such as medical information. Thus it is important to design statistical tests that guarantee the privacy of subjects in the data. In this work, we study hypothesis testing subject to differential privacy, specifically chi-squared tests for goodness of fit for multinomial data and independence between two categorical variables.
****
Pliable Rejection Sampling
Akram Erraqabi, Michal Valko, Alexandra Carpentier, Odalric Maillard
Rejection sampling is a technique for sampling from difficult distributions. However, its use is limited due to a high rejection rate. Common adaptive rejection sampling methods either work only for very specific distributions or without performance guarantees. In this paper, we present pliable rejection sampling (PRS), a new approach to rejection sampling, where we learn the sampling proposal using a kernel estimator. Since our method builds on rejection sampling, the samples obtained are with high probability i.i.d. and distributed according to f. Moreover, PRS comes with a guarantee on the number of accepted samples.
****
Differentially Private Policy Evaluation
Borja Balle, Maziar Gomrokchi, Doina Precup

We present the first differentially private algorithms for reinforcement learnin g, which apply to the task of evaluating a fixed policy. We establish two approa ches for achieving differential privacy, provide a theoretical analysis of the p rivacy and utility of the two algorithms, and show promising results on simple e mpirical examples.
****
Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning
Philip Thomas, Emma Brunskill
In this paper we present a new way of predicting the performance of a reinforcem ent learning policy given historical data that may have been generated by a diff erent policy. The ability to evaluate a policy from historical data is important for applications where the deployment of a bad policy can be dangerous or costl y. We show empirically that our algorithm produces estimates that often have ord ers of magnitude lower mean squared error than existing methods—it makes more ef ficient use of the available data. Our new estimator is based on two advances: a n extension of the doubly robust estimator (Jiang & Li, 2015), and a new way to mix between model based and importance sampling based estimates.
****
Discrete Deep Feature Extraction: A Theory and New Architectures
Thomas Wiatowski, Michael Tschannen, Aleksandar Stanic, Philipp Grohs, Helmut Bo elcskei
First steps towards a mathematical theory of deep convolutional neural networks for feature extraction were made—for the continuous-time case—in Mallat, 2012, a nd Wiatowski and Bölcskei, 2015. This paper considers the discrete case, introdu ces new convolutional neural network architectures, and proposes a mathematical framework for their analysis. Specifically, we establish deformation and transla tion sensitivity results of local and global nature, and we investigate how cert ain structural properties of the input signal are reflected in the corresponding feature vectors. Our theory applies to general filters and general Lipschitz-co ntinuous non-linearities and pooling operators. Experiments on handwritten digit classification and facial landmark detection—including feature importance evalu ation—complement the theoretical findings.
****
Efficient Algorithms for Adversarial Contextual Learning
Vasilis Syrgkanis, Akshay Krishnamurthy, Robert Schapire
We provide the first oracle efficient sublinear regret algorithms for adversaria l versions of the contextual bandit problem. In this problem, the learner repeat edly makes an action on the basis of a context and receives reward for the chose n action, with the goal of achieving reward competitive with a large class of po licies. We analyze two settings: i) in the transductive setting the learner know s the set of contexts a priori, ii) in the small separator setting, there exists a small set of contexts such that any two policies behave differently on one of the contexts in the set. Our algorithms fall into the Follow-The-Perturbed-Lead er family (Kalai and Vempala, 2005) and achieve regret O(T^3/4\sqrtK\log(N)) in the transductive setting and O(T^2/3 d^3/4 K\sqrt\log(N)) in the separator setti ng, where T is the number of rounds, K is the number of actions, N is the number of baseline policies, and d is the size of the separator. We actually solve the more general adversarial contextual semi-bandit linear optimization problem, wh ilst in the full information setting we address the even more general contextual combinatorial optimization. We provide several extensions and implications of o ur algorithms, such as switching regret and efficient learning with predictable sequences.
****
Training Deep Neural Networks via Direct Loss Minimization
Yang Song, Alexander Schwing, Richard, Raquel Urtasun
Supervised training of deep neural nets typically relies on minimizing cross-ent ropy. However, in many domains, we are interested in performing well on metrics specific to the application. In this paper we propose a direct loss minimization approach to train deep neural networks, which provably minimizes the applicatio n-specific loss function. This is often non-trivial, since these functions are n

either smooth nor decomposable and thus are not amenable to optimization with st
andard gradient-based methods. We demonstrate the effectiveness of our approach
in the context of maximizing average precision for ranking problems. Towards thi
s goal, we develop a novel dynamic programming algorithm that can efficiently co
mpute the weight updates. Our approach proves superior to a variety of baselines
 in the context of action classification and object detection, especially in the
 presence of label noise.
****
Sequence to Sequence Training of CTC-RNNs with Partial Windowing
Kyuyeon Hwang, Wonyong Sung
Connectionist temporal classification (CTC) based supervised sequence training o
f recurrent neural networks (RNNs) has shown great success in many machine learn
ing areas including end-to-end speech and handwritten character recognition. For
 the CTC training, however, it is required to unroll (or unfold) the RNN by the
length of an input sequence. This unrolling requires a lot of memory and hinders
 a small footprint implementation of online learning or adaptation. Furthermore,
 the length of training sequences is usually not uniform, which makes parallel t
raining with multiple sequences inefficient on shared memory models such as grap
hics processing units (GPUs). In this work, we introduce an expectation-maximiza
tion (EM) based online CTC algorithm that enables unidirectional RNNs to learn s
equences that are longer than the amount of unrolling. The RNNs can also be trai
ned to process an infinitely long input sequence without pre-segmentation or ext
ernal reset. Moreover, the proposed approach allows efficient parallel training
on GPUs. Our approach achieves 20.7% phoneme error rate (PER) on the very long i
nput sequence that is generated by concatenating all 192 utterances in the TIMIT
 core test set. In the end-to-end speech recognition task on the Wall Street Jou
rnal corpus, a network can be trained with only 64 times of unrolling with littl
e performance loss.
****
Variational Inference for Monte Carlo Objectives
Andriy Mnih, Danilo Rezende
Recent progress in deep latent variable models has largely been driven by the de
velopment of flexible and scalable variational inference methods. Variational tr
aining of this type involves maximizing a lower bound on the log-likelihood, usi
ng samples from the variational posterior to compute the required gradients. Rec
ently, Burda et al. (2016) have derived a tighter lower bound using a multi-samp
le importance sampling estimate of the likelihood and showed that optimizing it
yields models that use more of their capacity and achieve higher likelihoods. Th
is development showed the importance of such multi-sample objectives and explain
ed the success of several related approaches. We extend the multi-sample approac
h to discrete latent variables and analyze the difficulty encountered when estim
ating the gradients involved. We then develop the first unbiased gradient estima
tor designed for importance-sampled objectives and evaluate it at training gener
ative and structured output prediction models. The resulting estimator, which is
 based on low-variance per-sample learning signals, is both simpler and more eff
ective than the NVIL estimator proposed for the single-sample variational object
ive, and is competitive with the currently used biased estimators.
****
Hierarchical Decision Making In Electricity Grid Management
Gal Dalal, Elad Gilboa, Shie Mannor
The power grid is a complex and vital system that necessitates careful reliabili
ty management. Managing the grid is a difficult problem with multiple time scale
s of decision making and stochastic behavior due to renewable energy generations
, variable demand and unplanned outages. Solving this problem in the face of unc
ertainty requires a new methodology with tractable algorithms. In this work, we
introduce a new model for hierarchical decision making in complex systems. We ap
ply reinforcement learning (RL) methods to learn a proxy, i.e., a level of abstr
action, for real-time power grid reliability. We devise an algorithm that altern
ates between slow time-scale policy improvement, and fast time-scale value funct
ion approximation. We compare our results to prevailing heuristics, and show the

strength of our method.
****
Learning Sparse Combinatorial Representations via Two-stage Submodular Maximizat
ion
Eric Balkanski, Baharan Mirzasoleiman, Andreas Krause, Yaron Singer
We consider the problem of learning sparse representations of data sets, where t
he goal is to reduce a data set in manner that optimizes multiple objectives. Mo
tivated by applications of data summarization, we develop a new model which we r
efer to as the two-stage submodular maximization problem. This task can be viewe
d as a combinatorial analogue of representation learning problems such as dictio
nary learning and sparse regression. The two-stage problem strictly generalizes
the problem of cardinality constrained submodular maximization, though the objec
tive function is not submodular and the techniques for submodular maximization c
annot be applied. We describe a continuous optimization method which achieves an
approximation ratio which asymptotically approaches 1-1/e. For instances where
the asymptotics do not kick in, we design a local-search algorithm whose approxi
mation ratio is arbitrarily close to 1/2. We empirically demonstrate the effecti
veness of our methods on two multi-objective data summarization tasks, where the
goal is to construct summaries via sparse representative subsets w.r.t. to pred
efined objectives.
****
Understanding and Improving Convolutional Neural Networks via Concatenated Recti
fied Linear Units
Wenling Shang, Kihyuk Sohn, Diogo Almeida, Honglak Lee
Recently, convolutional neural networks (CNNs) have been used as a powerful tool
to solve many problems of machine learning and computer vision. In this paper,
we aim to provide insight on the property of convolutional neural networks, as w
ell as a generic method to improve the performance of many CNN architectures. Sp
ecifically, we first examine existing CNN models and observe an intriguing prope
rty that the filters in the lower layers form pairs (i.e., filters with opposite
phase). Inspired by our observation, we propose a novel, simple yet effective a
ctivation scheme called concatenated ReLU (CReLU) and theoretically analyze its
reconstruction property in CNNs. We integrate CReLU into several state-of-the-ar
t CNN architectures and demonstrate improvement in their recognition performance
on CIFAR-10/100 and ImageNet datasets with fewer trainable parameters. Our resu
lts suggest that better understanding of the properties of CNNs can lead to sign
ificant performance improvement with a simple modification.
****
Isotonic Hawkes Processes
Yichen Wang, Bo Xie, Nan Du, Le Song
Hawkes processes are powerful tools for modeling the mutual-excitation phenomena
commonly observed in event data from a variety of domains, such as social netwo
rks, quantitative finance and healthcare records. The intensity function of a Ha
wkes process is typically assumed to be linear in the sum of triggering kernels,
rendering it inadequate to capture nonlinear effects present in real-world data
. To address this shortcoming, we propose an Isotonic-Hawkes process whose inten
sity function is modulated by an additional nonlinear link function. We also dev
eloped a novel iterative algorithm which learns both the nonlinear link function
and other parameters provably. We showed that Isotonic-Hawkes processes can fit
a variety of nonlinear patterns which cannot be captured by conventional Hawkes
processes, and achieve superior empirical performance in real world application
s.
****
Cross-Graph Learning of Multi-Relational Associations
Hanxiao Liu, Yiming Yang
Cross-graph Relational Learning (CGRL) refers to the problem of predicting the s
trengths or labels of multi-relational tuples of heterogeneous object types, thr
ough the joint inference over multiple graphs which specify the internal connect
ions among each type of objects. CGRL is an open challenge in machine learning d
ue to the daunting number of all possible tuples to deal with when the numbers o

f nodes in multiple graphs are large, and because the labeled training instances are extremely sparse as typical. Existing methods such as tensor factorization or tensor-kernel machines do not work well because of the lack of convex formulation for the optimization of CGRL models, the poor scalability of the algorithms in handling combinatorial numbers of tuples, and/or the non-transductive nature of the learning methods which limits their ability to leverage unlabeled data in training. This paper proposes a novel framework which formulates CGRL as a convex optimization problem, enables transductive learning using both labeled and unlabeled tuples, and offers a scalable algorithm that guarantees the optimal solution and enjoys a constant time complexity with respect to the sizes of input graphs. In our experiments with a subset of DBLP publication records and an Enzyme multi-source dataset, the proposed method successfully scaled to the large cross-graph inference problem, and outperformed other representative approaches significantly.

****

Markov-modulated Marked Poisson Processes for Check-in Data

Jiangwei Pan, Vinayak Rao, Pankaj Agarwal, Alan Gelfand

We develop continuous-time probabilistic models to study trajectory data consisting of times and locations of user “check-ins”. We model the data as realizations of a marked point process, with intensity and mark-distribution modulated by a latent Markov jump process (MJP). We also include user-heterogeneity in our model by assigning each user a vector of “preferred locations”. Our model extends latent Dirichlet allocation by dropping the bag-of-words assumption and operating in continuous time. We show how an appropriate choice of priors allows efficient posterior inference. Our experiments demonstrate the usefulness of our approach by comparing with various baselines on a variety of tasks.

****

Beyond Parity Constraints: Fourier Analysis of Hash Functions for Inference

Tudor Achim, Ashish Sabharwal, Stefano Ermon

Random projections have played an important role in scaling up machine learning and data mining algorithms. Recently they have also been applied to probabilistic inference to estimate properties of high-dimensional distributions; however, they all rely on the same class of projections based on universal hashing. We provide a general framework to analyze random projections which relates their statistical properties to their Fourier spectrum, which is a well-studied area of theoretical computer science. Using this framework we introduce two new classes of hash functions for probabilistic inference and model counting that show promising performance on synthetic and real-world benchmarks.

****

On the Power and Limits of Distance-Based Learning

Periklis Papakonstantinou, Jia Xu, Guang Yang

We initiate the study of low-distortion finite metric embeddings in multi-class (and multi-label) classification where (i) both the space of input instances and the space of output classes have combinatorial metric structure and (ii) the concepts we wish to learn are low-distortion embeddings. We develop new geometric techniques and prove strong learning lower bounds. These provable limits hold even when we allow learners and classifiers to get advice by one or more experts. Our study overwhelmingly indicates that post-geometry assumptions are necessary in multi-class classification, as in natural language processing (NLP). Technically, the mathematical tools we developed in this work could be of independent interest to NLP. To the best of our knowledge, this is the first work which formally studies classification problems in combinatorial spaces. and where the concepts are low-distortion embeddings.

****

A Convex Atomic-Norm Approach to Multiple Sequence Alignment and Motif Discovery

Ian En-Hsu Yen, Xin Lin, Jiong Zhang, Pradeep Ravikumar, Inderjit Dhillon

Multiple Sequence Alignment and Motif Discovery, known as NP-hard problems, are two fundamental tasks in Bioinformatics. Existing approaches to these two problems are based on either local search methods such as Expectation Maximization (EM), Gibbs Sampling or greedy heuristic methods. In this work, we develop a convex

relaxation approach to both problems based on the recent concept of atomic norm and develop a new algorithm, termed Greedy Direction Method of Multiplier, for solving the convex relaxation with two convex atomic constraints. Experiments show that our convex relaxation approach produces solutions of higher quality than those standard tools widely-used in Bioinformatics community on the Multiple Sequence Alignment and Motif Discovery problems.
****

Generalized Direct Change Estimation in Ising Model Structure
Farideh Fazayeli, Arindam Banerjee
We consider the problem of estimating change in the dependency structure of two p-dimensional Ising models, based on respectively n_1 and n_2 samples drawn from the models. The change is assumed to be structured, e.g., sparse, block sparse, node-perturbed sparse, etc., such that it can be characterized by a suitable (atomic) norm. We present and analyze a norm-regularized estimator for directly estimating the change in structure, without having to estimate the structures of the individual Ising models. The estimator can work with any norm, and can be generalized to other graphical models under mild assumptions. We show that only one set of samples, say n_2, needs to satisfy the sample complexity requirement for the estimator to work, and the estimation error decreases as \fracc\sqrt\min(n_1,n_2), where c depends on the Gaussian width of the unit norm ball. For example, for \ell_1 norm applied to s-sparse change, the change can be accurately estimated with \min(n_1,n_2)=O(s \log p) which is sharper than an existing result n_1 = O(s^2 \log p) and n_2 = O(n_1^2). Experimental results illustrating the effectiveness of the proposed estimator are presented.
****

Robust Principal Component Analysis with Side Information
Kai-Yang Chiang, Cho-Jui Hsieh, Inderjit Dhillon
The robust principal component analysis (robust PCA) problem has been considered in many machine learning applications, where the goal is to decompose the data matrix as a low rank part plus a sparse residual. While current approaches are developed by only considering the low rank plus sparse structure, in many applications, side information of row and/or column entities may also be given, and it is still unclear to what extent could such information help robust PCA. Thus, in this paper, we study the problem of robust PCA with side information, where both prior structure and features of entities are exploited for recovery. We propose a convex problem to incorporate side information in robust PCA and show that the low rank matrix can be exactly recovered via the proposed method under certain conditions. In particular, our guarantee suggests that a substantial amount of low rank matrices, which cannot be recovered by standard robust PCA, become recoverable by our proposed method. The result theoretically justifies the effectiveness of features in robust PCA. In addition, we conduct synthetic experiments as well as a real application on noisy image classification to show that our method also improves the performance in practice by exploiting side information.
****

Towards Faster Rates and Oracle Property for Low-Rank Matrix Estimation
Huan Gui, Jiawei Han, Quanquan Gu
We present a unified framework for low-rank matrix estimation with a nonconvex penalty. A proximal gradient homotopy algorithm is proposed to solve the proposed optimization problem. Theoretically, we first prove that the proposed estimator attains a faster statistical rate than the traditional low-rank matrix estimator with nuclear norm penalty. Moreover, we rigorously show that under a certain condition on the magnitude of the nonzero singular values, the proposed estimator enjoys oracle property (i.e., exactly recovers the true rank of the matrix), besides attaining a faster rate. Extensive numerical experiments on both synthetic and real world datasets corroborate our theoretical findings.
****

Early and Reliable Event Detection Using Proximity Space Representation
Maxime Sangnier, Jerome Gauthier, Alain Rakotomamonjy
Let us consider a specific action or situation (called event) that takes place within a time series. The objective in early detection is to build a decision fun

ction that is able to go off as soon as possible from the onset of an occurrence of this event. This implies making a decision with an incomplete information. This paper proposes a novel framework that i) guarantees that a detection made with a partial observation will also occur at full observation of the time-series; ii) incorporates in a consistent manner the lack of knowledge about the minimal amount of information needed to make a decision. The proposed detector is based on mapping the temporal sequences to a landmarking space thanks to appropriately designed similarity functions. As a by-product, the framework benefits from a scalable training algorithm and a theoretical guarantee concerning its generalization ability. We also discuss an important improvement of our framework in which decision function can still be made reliable while being more expressive. Our experimental studies provide compelling results on toy data, presenting the trade-off that occurs when aiming at accuracy, earliness and reliability. Results on real physiological and video datasets show that our proposed approach is as accurate and early as state-of-the-art algorithm, while ensuring reliability and being far more efficient to learn.

****

Stratified Sampling Meets Machine Learning

Edo Liberty, Kevin Lang, Konstantin Shmakov

This paper solves a specialized regression problem to obtain sampling probabilities for records in databases. The goal is to sample a small set of records over which evaluating aggregate queries can be done both efficiently and accurately. We provide a principled and provable solution for this problem; it is parameterless and requires no data insights. Unlike standard regression problems, the loss is inversely proportional to the regressed-to values. Moreover, a cost zero solution always exists and can only be excluded by hard budget constraints. A unique form of regularization is also needed. We provide an efficient and simple regularized Empirical Risk Minimization (ERM) algorithm along with a theoretical generalization result. Our extensive experimental results significantly improve over both uniform sampling and standard stratified sampling which are de-facto the industry standards.

****

Efficient Multi-Instance Learning for Activity Recognition from Time Series Data Using an Auto-Regressive Hidden Markov Model

Xinze Guan, Raviv Raich, Weng-Keen Wong

Activity recognition from sensor data has spurred a great deal of interest due to its impact on health care. Prior work on activity recognition from multivariate time series data has mainly applied supervised learning techniques which require a high degree of annotation effort to produce training data with the start and end times of each activity. In order to reduce the annotation effort, we present a weakly supervised approach based on multi-instance learning. We introduce a generative graphical model for multi-instance learning on time series data based on an auto-regressive hidden Markov model. Our model has a number of advantages, including the ability to produce both bag and instance-level predictions as well as an efficient exact inference algorithm based on dynamic programming.

****

Generalization Properties and Implicit Regularization for Multiple Passes SGM

Junhong Lin, Raffaello Camoriano, Lorenzo Rosasco

We study the generalization properties of stochastic gradient methods for learning with convex loss functions and linearly parameterized functions. We show that, in the absence of penalizations or constraints, the stability and approximation properties of the algorithm can be controlled by tuning either the step-size or the number of passes over the data. In this view, these parameters can be seen to control a form of implicit regularization. Numerical results complement the theoretical findings.

****

Principal Component Projection Without Principal Component Analysis

Roy Frostig, Cameron Musco, Christopher Musco, Aaron Sidford

We show how to efficiently project a vector onto the top principal components of a matrix, *without explicitly computing these components*. Specifically, we int

roduce an iterative algorithm that provably computes the projection using few ca lls to any black-box routine for ridge regression. By avoiding explicit principa l component analysis (PCA), our algorithm is the first with no runtime dependenc e on the number of top principal components. We show that it can be used to give a fast iterative method for the popular principal component regression problem, giving the first major runtime improvement over the naive method of combining P CA with regression. To achieve our results, we first observe that ridge regressi on can be used to obtain a "smooth projection" onto the top principal components . We then sharpen this approximation to true projection using a low-degree polyn omial approximation to the matrix step function. Step function approximation is a topic of long-term interest in scientific computing. We extend prior theory by constructing polynomials with simple iterative structure and rigorously analyzi ng their behavior under limited precision.

****

Recovery guarantee of weighted low-rank approximation via alternating minimizati on

Yuanzhi Li, Yingyu Liang, Andrej Risteski

Many applications require recovering a ground truth low-rank matrix from noisy o bservations of the entries, which in practice is typically formulated as a weigh ted low-rank approximation problem and solved by non-convex optimization heurist ics such as alternating minimization. In this paper, we provide provable recover y guarantee of weighted low-rank via a simple alternating minimization algorithm . In particular, for a natural class of matrices and weights and without any ass umption on the noise, we bound the spectral norm of the difference between the r ecovered matrix and the ground truth, by the spectral norm of the weighted noise plus an additive error term that decreases exponentially with the number of rou nds of alternating minimization, from either initialization by SVD or, more impo rtantly, random initialization. These provide the first theoretical results for weighted low-rank approximation via alternating minimization with non-binary det erministic weights, significantly generalizing those for matrix completion, the special case with binary weights, since our assumptions are similar or weaker th an those made in existing works. Furthermore, this is achieved by a very simple algorithm that improves the vanilla alternating minimization with a simple clipp ing step.

****

Deconstructing the Ladder Network Architecture

Mohammad Pezeshki, Linxi Fan, Philemon Brakel, Aaron Courville, Yoshua Bengio

The Ladder Network is a recent new approach to semi-supervised learning that tur ned out to be very successful. While showing impressive performance, the Ladder Network has many components intertwined, whose contributions are not obvious in such a complex architecture. This paper presents an extensive experimental inves tigation of variants of the Ladder Network in which we replaced or removed indiv idual components to learn about their relative importance. For semi-supervised t asks, we conclude that the most important contribution is made by the lateral co nnections, followed by the application of noise, and the choice of what we refer to as the ‘combinator function’. As the number of labeled training examples inc reases, the lateral connections and the reconstruction criterion become less imp ortant, with most of the generalization improvement coming from the injection of noise in each layer. Finally, we introduce a combinator function that reduces t est error rates on Permutation-Invariant MNIST to 0.57% for the supervised setti ng, and to 0.97% and 1.0% for semi-supervised settings with 1000 and 100 labeled examples, respectively.

****

Generalization and Exploration via Randomized Value Functions

Ian Osband, Benjamin Van Roy, Zheng Wen

We propose randomized least-squares value iteration (RLSVI) – a new reinforcemen t learning algorithm designed to explore and generalize efficiently via linearly parameterized value functions. We explain why versions of least-squares value i teration that use Boltzmann or epsilon-greedy exploration can be highly ineffici ent, and we present computational results that demonstrate dramatic efficiency g

ains enjoyed by RLSVI. Further, we establish an upper bound on the expected regr
et of RLSVI that demonstrates near-optimality in a tabula rasa learning context.
 More broadly, our results suggest that randomized value functions offer a promi
sing approach to tackling a critical challenge in reinforcement learning: synthe
sizing efficient exploration and effective generalization.
****

Evasion and Hardening of Tree Ensemble Classifiers
Alex Kantchelian, J. D. Tygar, Anthony Joseph
Classifier evasion consists in finding for a given instance x the “nearest” inst
ance x′ such that the classifier predictions of x and x′ are different. We prese
nt two novel algorithms for systematically computing evasions for tree ensembles
 such as boosted trees and random forests. Our first algorithm uses a Mixed Inte
ger Linear Program solver and finds the optimal evading instance under an expres
sive set of constraints. Our second algorithm trades off optimality for speed by
 using symbolic prediction, a novel algorithm for fast finite differences on tre
e ensembles. On a digit recognition task, we demonstrate that both gradient boos
ted trees and random forests are extremely susceptible to evasions. Finally, we
harden a boosted tree model without loss of predictive accuracy by augmenting th
e training set of each boosting round with evading instances, a technique we cal
l adversarial boosting.
****

Dynamic Memory Networks for Visual and Textual Question Answering
Caiming Xiong, Stephen Merity, Richard Socher
Neural network architectures with memory and attention mechanisms exhibit certai
n reason- ing capabilities required for question answering. One such architectur
e, the dynamic memory net- work (DMN), obtained high accuracy on a variety of la
nguage tasks. However, it was not shown whether the architecture achieves strong
 results for question answering when supporting facts are not marked during trai
ning or whether it could be applied to other modalities such as images. Based on
 an analysis of the DMN, we propose several improvements to its memory and input
 modules. Together with these changes we introduce a novel input module for imag
es in order to be able to answer visual questions. Our new DMN+ model improves t
he state of the art on both the Visual Question Answering dataset and the bAbI-1
0k text question-answering dataset without supporting fact supervision.
****

Estimating Cosmological Parameters from the Dark Matter Distribution
Siamak Ravanbakhsh, Junier Oliva, Sebastian Fromenteau, Layne Price, Shirley Ho,
 Jeff Schneider, Barnabas Poczos
A grand challenge of the 21st century cosmology is to accurately estimate the co
smological parameters of our Universe. A major approach in estimating the cosmol
ogical parameters is to use the large scale matter distribution of the Universe.
 Galaxy surveys provide the means to map out cosmic large-scale structure in thr
ee dimensions. Information about galaxy locations is typically summarized in a "
single" function of scale, such as the galaxy correlation function or power-spec
trum. We show that it is possible to estimate these cosmological parameters dire
ctly from the distribution of matter. This paper presents the application of dee
p 3D convolutional networks to volumetric representation of dark matter simulati
ons as well as the results obtained using a recently proposed distribution regre
ssion framework, showing that machine learning techniques are comparable to, and
 can sometimes outperform, maximum-likelihood point estimates using "cosmologica
l models". This opens the way to estimating the parameters of our Universe with
higher accuracy.
****

Learning Population-Level Diffusions with Generative RNNs
Tatsunori Hashimoto, David Gifford, Tommi Jaakkola
We estimate stochastic processes that govern the dynamics of evolving population
s such as cell differentiation. The problem is challenging since longitudinal tr
ajectory measurements of individuals in a population are rarely available due to
 experimental cost and/or privacy. We show that cross-sectional samples from an
evolving population suffice for recovery within a class of processes even if sam

ples are available only at a few distinct time points. We provide a stratified a nalysis of recoverability conditions, and establish that reversibility is suffic ient for recoverability. For estimation, we derive a natural loss and regulariza tion, and parameterize the processes as diffusive recurrent neural networks. We demonstrate the approach in the context of uncovering complex cellular dynamics known as the ‘epigenetic landscape’ from existing biological assays.
****

Expressiveness of Rectifier Networks
Xingyuan Pan, Vivek Srikumar
Rectified Linear Units (ReLUs) have been shown to ameliorate the vanishing gradi ent problem, allow for efficient backpropagation, and empirically promote sparsi ty in the learned parameters. They have led to state-of-the-art results in a var iety of applications. However, unlike threshold and sigmoid networks, ReLU netwo rks are less explored from the perspective of their expressiveness. This paper s tudies the expressiveness of ReLU networks. We characterize the decision boundar y of two-layer ReLU networks by constructing functionally equivalent threshold n etworks. We show that while the decision boundary of a two-layer ReLU network ca n be captured by a threshold network, the latter may require an exponentially la rger number of hidden units. We also formulate sufficient conditions for a corre sponding logarithmic reduction in the number of hidden units to represent a sign network as a ReLU network. Finally, we experimentally compare threshold network s and their much smaller ReLU counterparts with respect to their ability to lear n from synthetically generated data.
****

Discrete Distribution Estimation under Local Privacy
Peter Kairouz, Keith Bonawitz, Daniel Ramage
The collection and analysis of user data drives improvements in the app and web ecosystems, but comes with risks to privacy. This paper examines discrete distri bution estimation under local privacy, a setting wherein service providers can l earn the distribution of a categorical statistic of interest without collecting the underlying data. We present new mechanisms, including hashed k-ary Randomize d Response (KRR), that empirically meet or exceed the utility of existing mechan isms at all privacy levels. New theoretical results demonstrate the order-optima lity of KRR and the existing RAPPOR mechanism at different privacy regimes.
****

Square Root Graphical Models: Multivariate Generalizations of Univariate Exponen tial Families that Permit Positive Dependencies
David Inouye, Pradeep Ravikumar, Inderjit Dhillon
We develop Square Root Graphical Models (SQR), a novel class of parametric graph ical models that provides multivariate generalizations of univariate exponential family distributions. Previous multivariate graphical models [Yang et al. 2015] did not allow positive dependencies for the exponential and Poisson generalizat ions. However, in many real-world datasets, variables clearly have positive depe ndencies. For example, the airport delay time in New York—modeled as an exponent ial distribution—is positively related to the delay time in Boston. With this mo tivation, we give an example of our model class derived from the univariate expo nential distribution that allows for almost arbitrary positive and negative depe ndencies with only a mild condition on the parameter matrix—a condition akin to the positive definiteness of the Gaussian covariance matrix. Our Poisson general ization allows for both positive and negative dependencies without any constrain ts on the parameter values. We also develop parameter estimation methods using n ode-wise regressions with \ell_1 regularization and likelihood approximation met hods using sampling. Finally, we demonstrate our exponential generalization on a synthetic dataset and a real-world dataset of airport delay times.
****

A Box-Constrained Approach for Hard Permutation Problems
Cong Han Lim, Steve Wright
We describe the use of sorting networks to form relaxations of problems involvin g permutations of n objects. This approach is an alternative to relaxations base d on the Birkhoff polytope (the set of n \times n doubly stochastic matrices), p

roviding a more compact formulation in which the only constraints are box constr aints. Using this approach, we form a variant of the relaxation of the quadratic assignment problem recently studied in Vogelstein et al. (2015), and show that the continuation method applied to this formulation can be quite effective. We d evelop a coordinate descent algorithm that achieves a per-cycle complexity of O(n^2 \log^2 n). We compare this method with Fast Approximate QAP (FAQ) algorithm introduced in Vogelstein et al. (2015), which uses a conditional-gradient method whose per-iteration complexity is O(n^3). We demonstrate that for most problems in QAPLIB and for a class of synthetic QAP problems, the sorting-network formul ation returns solutions that are competitive with the FAQ algorithm, often in si gnificantly less computing time.

****

Geometric Mean Metric Learning

Pourya Zadeh, Reshad Hosseini, Suvrit Sra

We revisit the task of learning a Euclidean metric from data. We approach this p roblem from first principles and formulate it as a surprisingly simple optimizat ion problem. Indeed, our formulation even admits a closed form solution. This so lution possesses several very attractive properties: (i) an innate geometric app eal through the Riemannian geometry of positive definite matrices; (ii) ease of interpretability; and (iii) computational speed several orders of magnitude fast er than the widely used LMNN and ITML methods. Furthermore, on standard benchmar k datasets, our closed-form solution consistently attains higher classification accuracy.

****

Sparse Nonlinear Regression: Parameter Estimation under Nonconvexity

Zhuoran Yang, Zhaoran Wang, Han Liu, Yonina Eldar, Tong Zhang

We study parameter estimation for sparse nonlinear regression. More specifically , we assume the data are given by y = f(\bf x^T \bf β^*) + ε, where f is nonli near. To recover \bf βs, we propose an \ell_1-regularized least-squares estimato r. Unlike classical linear regression, the corresponding optimization problem is nonconvex because of the nonlinearity of f. In spite of the nonconvexity, we pr ove that under mild conditions, every stationary point of the objective enjoys a n optimal statistical rate of convergence. Detailed numerical results are provid ed to back up our theory.

****

Conditional Bernoulli Mixtures for Multi-label Classification

Cheng Li, Bingyu Wang, Virgil Pavlu, Javed Aslam

Multi-label classification is an important machine learning task wherein one ass igns a subset of candidate labels to an object. In this paper, we propose a new multi-label classification method based on Conditional Bernoulli Mixtures. Our p roposed method has several attractive properties: it captures label dependencies ; it reduces the multi-label problem to several standard binary and multi-class problems; it subsumes the classic independent binary prediction and power-set su bset prediction methods as special cases; and it exhibits accuracy and/or comput ational complexity advantages over existing approaches. We demonstrate two imple mentations of our method using logistic regressions and gradient boosted trees, together with a simple training procedure based on Expectation Maximization. We further derive an efficient prediction procedure based on dynamic programming, t hus avoiding the cost of examining an exponential number of potential label subs ets. Experimental results show the effectiveness of the proposed method against competitive alternatives on benchmark datasets.

****

Scalable Discrete Sampling as a Multi-Armed Bandit Problem

Yutian Chen, Zoubin Ghahramani

Drawing a sample from a discrete distribution is one of the building components for Monte Carlo methods. Like other sampling algorithms, discrete sampling suffe rs from the high computational burden in large-scale inference problems. We stud y the problem of sampling a discrete random variable with a high degree of depen dency that is typical in large-scale Bayesian inference and graphical models, an d propose an efficient approximate solution with a subsampling approach. We make

a novel connection between the discrete sampling and Multi-Armed Bandits proble
ms with a finite reward population and provide three algorithms with theoretical
guarantees. Empirical evaluations show the robustness and efficiency of the app
roximate algorithms in both synthetic and real-world large-scale problems.
****

Recycling Randomness with Structure for Sublinear time Kernel Expansions
Krzysztof Choromanski, Vikas Sindhwani
We propose a scheme for recycling Gaussian random vectors into structured matric
es to ap- proximate various kernel functions in sublin- ear time via random embe
ddings. Our frame- work includes the Fastfood construction of Le et al. (2013) a
s a special case, but also ex- tends to Circulant, Toeplitz and Hankel matri- ce
s, and the broader family of structured matri- ces that are characterized by the
concept of low- displacement rank. We introduce notions of co- herence and grap
h-theoretic structural constants that control the approximation quality, and pro
ve unbiasedness and low-variance properties of ran- dom feature maps that arise
within our frame- work. For the case of low-displacement matri- ces, we show how
the degree of structure and randomness can be controlled to reduce statis- tica
l variance at the cost of increased computa- tion and storage requirements. Empi
rical results strongly support our theory and justify the use of a broader famil
y of structured matrices for scal- ing up kernel methods using random features.
****

Bidirectional Helmholtz Machines
Jorg Bornschein, Samira Shabanian, Asja Fischer, Yoshua Bengio
Efficient unsupervised training and inference in deep generative models remains
a challenging problem. One basic approach, called Helmholtz machine or Variation
al Autoencoder, involves training a top-down directed generative model together
with a bottom-up auxiliary model used for approximate inference. Recent results
indicate that better generative models can be obtained with better approximate i
nference procedures. Instead of improving the inference procedure, we here propo
se a new model, the bidirectional Helmholtz machine, which guarantees that the t
op-down and bottom-up distributions can efficiently invert each other. We achiev
e this by interpreting both the top-down and the bottom-up directed models as ap
proximate inference distributions and by defining the model distribution to be t
he geometric mean of these two. We present a lower-bound for the likelihood of t
his model and we show that optimizing this bound regularizes the model so that t
he Bhattacharyya distance between the bottom-up and top-down approximate distrib
utions is minimized. This approach results in state of the art generative models
which prefer significantly deeper architectures while it allows for orders of m
agnitude more efficient likelihood estimation.
****

Faster Convex Optimization: Simulated Annealing with an Efficient Universal Barr
ier
Jacob Abernethy, Elad Hazan
This paper explores a surprising equivalence between two seemingly-distinct conv
ex optimization methods. We show that simulated annealing, a well-studied random
walk algorithms, is *directly equivalent*, in a certain sense, to the central p
ath interior point algorithm for the the entropic universal barrier function. Th
is connection exhibits several benefits. First, we are able improve the state of
the art time complexity for convex optimization under the membership oracle mod
el by devising a new temperature schedule for simulated annealing motivated by c
entral path following interior point methods. Second, we get an efficient random
ized interior point method with an efficiently computable universal barrier for
any convex set described by a membership oracle. Previously, efficiently computa
ble barriers were known only for particular convex sets.
****

Preconditioning Kernel Matrices
Kurt Cutajar, Michael Osborne, John Cunningham, Maurizio Filippone
The computational and storage complexity of kernel machines presents the primary
barrier to their scaling to large, modern, datasets. A common way to tackle the
scalability issue is to use the conjugate gradient algorithm, which relieves th

e constraints on both storage (the kernel matrix need not be stored) and computa
tion (both stochastic gradients and parallelization can be used). Even so, conju
gate gradient is not without its own issues: the conditioning of kernel matrices
 is often such that conjugate gradients will have poor convergence in practice.
Preconditioning is a common approach to alleviating this issue. Here we propose
preconditioned conjugate gradients for kernel machines, and develop a broad rang
e of preconditioners particularly useful for kernel matrices. We describe a scal
able approach to both solving kernel machines and learning their hyperparameters
. We show this approach is exact in the limit of iterations and outperforms stat
e-of-the-art approximations for a given computational budget.
****

Greedy Column Subset Selection: New Bounds and Distributed Algorithms
Jason Altschuler, Aditya Bhaskara, Gang Fu, Vahab Mirrokni, Afshin Rostamizadeh,
 Morteza Zadimoghaddam
The problem of column subset selection has recently attracted a large body of re
search, with feature selection serving as one obvious and important application.
 Among the techniques that have been applied to solve this problem, the greedy a
lgorithm has been shown to be quite effective in practice. However, theoretical
guarantees on its performance have not been explored thoroughly, especially in a
 distributed setting. In this paper, we study the greedy algorithm for the colum
n subset selection problem from a theoretical and empirical perspective and show
 its effectiveness in a distributed setting. In particular, we provide an improv
ed approximation guarantee for the greedy algorithm which we show is tight up to
 a constant factor, and present the first distributed implementation with provab
le approximation factors. We use the idea of randomized composable core-sets, de
veloped recently in the context of submodular maximization. Finally, we validate
 the effectiveness of this distributed algorithm via an empirical study.
****

Dynamic Capacity Networks
Amjad Almahairi, Nicolas Ballas, Tim Cooijmans, Yin Zheng, Hugo Larochelle, Aaro
n Courville
We introduce the Dynamic Capacity Network (DCN), a neural network that can adapt
ively assign its capacity across different portions of the input data. This is a
chieved by combining modules of two types: low-capacity sub-networks and high-ca
pacity sub-networks. The low-capacity sub-networks are applied across most of th
e input, but also provide a guide to select a few portions of the input on which
 to apply the high-capacity sub-networks. The selection is made using a novel gr
adient-based attention mechanism, that efficiently identifies input regions for
which the DCN’s output is most sensitive and to which we should devote more capa
city. We focus our empirical evaluation on the Cluttered MNIST and SVHN image da
tasets. Our findings indicate that DCNs are able to drastically reduce the numbe
r of computations, compared to traditional convolutional neural networks, while
maintaining similar or even better performance.
****

Pricing a Low-regret Seller
Hoda Heidari, Mohammad Mahdian, Umar Syed, Sergei Vassilvitskii, Sadra Yazdanbod
As the number of ad exchanges has grown, publishers have turned to low regret le
arning algorithms to decide which exchange offers the best price for their inven
tory. This in turn opens the following question for the exchange: how to set pri
ces to attract as many sellers as possible and maximize revenue. In this work we
 formulate this precisely as a learning problem, and present algorithms showing
that by simply knowing that the counterparty is using a low regret algorithm is
enough for the exchange to have its own low regret learning algorithm to find th
e optimal price.
****

Estimation from Indirect Supervision with Linear Moments
Aditi Raghunathan, Roy Frostig, John Duchi, Percy Liang
In structured prediction problems where we have indirect supervision of the outp
ut, maximum marginal likelihood faces two computational obstacles: non-convexity
 of the objective and intractability of even a single gradient computation. In t

his paper, we bypass both obstacles for a class of what we call linear indirectly-supervised problems. Our approach is simple: we solve a linear system to estimate sufficient statistics of the model, which we then use to estimate parameters via convex optimization. We analyze the statistical properties of our approach and show empirically that it is effective in two settings: learning with local privacy constraints and learning from low-cost count-based annotations.
****

Speeding up k-means by approximating Euclidean distances via block vectors
Thomas Bottesch, Thomas Bühler, Markus Kächele
This paper introduces a new method to approximate Euclidean distances between points using block vectors in combination with the Hölder inequality. By defining lower bounds based on the proposed approximation, cluster algorithms can be considerably accelerated without loss of quality. In extensive experiments, we show a considerable reduction in terms of computational time in comparison to standard methods and the recently proposed Yinyang k-means. Additionally we show that the memory consumption of the presented clustering algorithm does not depend on the number of clusters, which makes the approach suitable for large scale problems.
****

Learning and Inference via Maximum Inner Product Search
Stephen Mussmann, Stefano Ermon
A large class of commonly used probabilistic models known as log-linear models are defined up to a normalization constant.Typical learning algorithms for such models require solving a sequence of probabilistic inference queries. These inferences are typically intractable, and are a major bottleneck for learning models with large output spaces. In this paper, we provide a new approach for amortizing the cost of a sequence of related inference queries, such as the ones arising during learning. Our technique relies on a surprising connection with algorithms developed in the past two decades for similarity search in large data bases. Our approach achieves improved running times with provable approximation guarantees. We show that it performs well both on synthetic data and neural language models with large output spaces.
****

A Superlinearly-Convergent Proximal Newton-type Method for the Optimization of Finite Sums
Anton Rodomanov, Dmitry Kropotov
We consider the problem of minimizing the strongly convex sum of a finite number of convex functions. Standard algorithms for solving this problem in the class of incremental/stochastic methods have at most a linear convergence rate. We propose a new incremental method whose convergence rate is superlinear – the Newton-type incremental method (NIM). The idea of the method is to introduce a model of the objective with the same sum-of-functions structure and further update a single component of the model per iteration. We prove that NIM has a superlinear local convergence rate and linear global convergence rate. Experiments show that the method is very effective for problems with a large number of functions and a small number of variables.
****

A Kernel Test of Goodness of Fit
Kacper Chwialkowski, Heiko Strathmann, Arthur Gretton
We propose a nonparametric statistical test for goodness-of-fit: given a set of samples, the test determines how likely it is that these were generated from a target density function. The measure of goodness-of-fit is a divergence constructed via Stein’s method using functions from a Reproducing Kernel Hilbert Space. Our test statistic is based on an empirical estimate of this divergence, taking the form of a V-statistic in terms of the log gradients of the target density and the kernel. We derive a statistical test, both for i.i.d. and non-i.i.d. samples, where we estimate the null distribution quantiles using a wild bootstrap procedure. We apply our test to quantifying convergence of approximate Markov Chain Monte Carlo methods, statistical model criticism, and evaluating quality of fit vs model complexity in nonparametric density estimation.

****
Interacting Particle Markov Chain Monte Carlo
Tom Rainforth, Christian Naesseth, Fredrik Lindsten, Brooks Paige, Jan-Willem Va ndemeent, Arnaud Doucet, Frank Wood
We introduce interacting particle Markov chain Monte Carlo (iPMCMC), a PMCMC met hod based on an interacting pool of standard and conditional sequential Monte Ca rlo samplers. Like related methods, iPMCMC is a Markov chain Monte Carlo sampler on an extended space. We present empirical results that show significant improv ements in mixing rates relative to both non-interacting PMCMC samplers and a sin gle PMCMC sampler with an equivalent memory and computational budget. An additio nal advantage of the iPMCMC method is that it is suitable for distributed and mu lti-core architectures.
****
Faster Eigenvector Computation via Shift-and-Invert Preconditioning
Dan Garber, Elad Hazan, Chi Jin, Sham, Cameron Musco, Praneeth Netrapalli, Aaro n Sidford
We give faster algorithms and improved sample complexities for the fundamental p roblem of estimating the top eigenvector. Given an explicit matrix $A \in \mathb b{R}^{n \times d}$, we show how to compute an ϵ-approximate top eigenve ctor of A^TA in time $\tilde O\left(\left[\text{nnz}(A) + \frac{d \text{sr}(A)}{\text{gap}^2} \right] \cdot \log 1/\epsilon\right)$. Here $\text{nnz}(A)$ is the number of nonzeros in A, $\text{sr}(A)$ is the stable rank, and gap is the relative eigengap. We also consider an online setting in which, given a stream of i.i.d. samples from a distribution D with covariance matrix Σ and a ve ctor x_0 which is an $O(\text{gap})$ approximate top eigenvector for Σ, we show how to refine x_0 to an ϵ approximation using $O \left(\fr ac{\text{var}(\mathcal{D})}{\text{gap}-\epsilon}\right)$ samples from $\mathcal{ D}$. Here $\text{var}(\mathcal{D})$ is a natural notion of variance. Combining o ur algorithm with previous work to initialize x_0, we obtain improved sample c omplexities and runtimes under a variety of assumptions on D. We achieve our res ults via a robust analysis of the classic shift-and-invert preconditioning metho d. This technique lets us reduce eigenvector computation to approximately solvin g a series of linear systems with fast stochastic gradient methods.
****
A Theory of Generative ConvNet
Jianwen Xie, Yang Lu, Song-Chun Zhu, Yingnian Wu
We show that a generative random field model, which we call generative ConvNet, can be derived from the commonly used discriminative ConvNet, by assuming a Conv Net for multi-category classification and assuming one of the category is a base category generated by a reference distribution. If we further assume that the n on-linearity in the ConvNet is Rectified Linear Unit (ReLU) and the reference di stribution is Gaussian white noise, then we obtain a generative ConvNet model th at is unique among energy-based models: The model is piecewise Gaussian, and the means of the Gaussian pieces are defined by an auto-encoder, where the filters in the bottom-up encoding become the basis functions in the top-down decoding, a nd the binary activation variables detected by the filters in the bottom-up conv olution process become the coefficients of the basis functions in the top-down d econvolution process. The Langevin dynamics for sampling the generative ConvNet is driven by the reconstruction error of this auto-encoder. The contrastive dive rgence learning of the generative ConvNet reconstructs the training images by th e auto-encoder. The maximum likelihood learning algorithm can synthesize realist ic natural image patterns.
****
Efficient Learning with a Family of Nonconvex Regularizers by Redistributing Non convexity
Quanming Yao, James Kwok
The use of convex regularizers allow for easy optimization, though they often pr oduce biased estimation and inferior prediction performance. Recently, nonconvex regularizers have attracted a lot of attention and outperformed convex ones. Ho wever, the resultant optimization problem is much harder. In this paper, for a l

arge class of nonconvex regularizers, we propose to move the nonconvexity from t
he regularizer to the loss. The nonconvex regularizer is then transformed to a f
amiliar convex regularizer, while the resultant loss function can still be guara
nteed to be smooth. Learning with the convexified regularizer can be performed b
y existing efficient algorithms originally designed for convex regularizers (suc
h as the standard proximal algorithm and Frank-Wolfe algorithm). Moreover, it ca
n be shown that critical points of the transformed problem are also critical poi
nts of the original problem. Extensive experiments on a number of nonconvex regu
larization problems show that the proposed procedure is much faster than the sta
te-of-the-art nonconvex solvers.
****

Computationally Efficient Nyström Approximation using Fast Transforms
Si Si, Cho-Jui Hsieh, Inderjit Dhillon
Our goal is to improve the \it training and \it prediction time of Nyström metho
d, which is a widely-used technique for generating low-rank kernel matrix approx
imations. When applying the Nyström approximation for large-scale applications,
both training and prediction time is dominated by computing kernel values betwee
n a data point and all landmark points. With m landmark points, this computation
 requires Θ(md) time (flops), where d is the input dimension. In this paper, we
propose the use of a family of fast transforms to generate structured landmark p
oints for Nyström approximation. By exploiting fast transforms, e.g., Haar trans
form and Hadamard transform, our modified Nyström method requires only Θ(m) or Θ
(m\log d) time to compute the kernel values between a given data point and m lan
dmark points. This improvement in time complexity can significantly speed up ker
nel approximation and benefit prediction speed in kernel machines. For instance,
 on the webspam data (more than 300,000 data points), our proposed algorithm ena
bles kernel SVM prediction to deliver 98% accuracy and the resulting prediction
time is 1000 times faster than LIBSVM and only 10 times slower than linear SVM p
rediction (which yields only 91% accuracy).
****

Gromov-Wasserstein Averaging of Kernel and Distance Matrices
Gabriel Peyré, Marco Cuturi, Justin Solomon
This paper presents a new technique for computing the barycenter of a set of dis
tance or kernel matrices. These matrices, which define the inter-relationships b
etween points sampled from individual domains, are not required to have the same
 size or to be in row-by-row correspondence. We compare these matrices using the
 softassign criterion, which measures the minimum distortion induced by a probab
ilistic map from the rows of one similarity matrix to the rows of another; this
criterion amounts to a regularized version of the Gromov-Wasserstein (GW) distan
ce between metric-measure spaces. The barycenter is then defined as a Fréchet me
an of the input matrices with respect to this criterion, minimizing a weighted s
um of softassign values. We provide a fast iterative algorithm for the resulting
 nonconvex optimization problem, built upon state-of- the-art tools for regulari
zed optimal transportation. We demonstrate its application to the computation of
 shape barycenters and to the prediction of energy levels from molecular configu
rations in quantum chemistry.
****

Robust Monte Carlo Sampling using Riemannian Nosé-Poincaré Hamiltonian Dynamics
Anirban Roychowdhury, Brian Kulis, Srinivasan Parthasarathy
We present a Monte Carlo sampler using a modified Nosé-Poincaré Hamiltonian alon
g with Riemannian preconditioning. Hamiltonian Monte Carlo samplers allow better
 exploration of the state space as opposed to random walk-based methods, but, fr
om a molecular dynamics perspective, may not necessarily provide samples from th
e canonical ensemble. Nosé-Hoover samplers rectify that shortcoming, but the res
ultant dynamics are not Hamiltonian. Furthermore, usage of these algorithms on l
arge real-life datasets necessitates the use of stochastic gradients, which acts
 as another potentially destabilizing source of noise. In this work, we propose
dynamics based on a modified Nosé-Poincaré Hamiltonian augmented with Riemannian
 manifold corrections. The resultant symplectic sampling algorithm samples from
the canonical ensemble while using structural cues from the Riemannian precondit

ioning matrices to efficiently traverse the parameter space. We also propose a s
tochastic variant using additional terms in the Hamiltonian to correct for the n
oise from the stochastic gradients. We show strong performance of our algorithms
 on synthetic datasets and high-dimensional Poisson factor analysis-based topic
modeling scenarios.
****

The Segmented iHMM: A Simple, Efficient Hierarchical Infinite HMM
Ardavan Saeedi, Matthew Hoffman, Matthew Johnson, Ryan Adams
We propose the segmented iHMM (siHMM), a hierarchical infinite hidden Markov mod
el (iHMM) that supports a simple, efficient inference scheme. The siHMM is well
suited to segmentation problems, where the goal is to identify points at which a
 time series transitions from one relatively stable regime to a new regime. Conv
entional iHMMs often struggle with such problems, since they have no mechanism f
or distinguishing between high-and low-level dynamics. Hierarchical HMMs (HHMMs)
 can do better, but they require much more complex and expensive inference algor
ithms. The siHMM retains the simplicity and efficiency of the iHMM, but outperfo
rms it on a variety of segmentation problems, achieving performance that matches
 or exceeds that of a more complicated HHMM.
****

Meta–Gradient Boosted Decision Tree Model for Weight and Target Learning
Yury Ustinovskiy, Valentina Fedorova, Gleb Gusev, Pavel Serdyukov
Labeled training data is an essential part of any supervised machine learning fr
amework. In practice, there is a trade-off between the quality of a label and it
s cost. In this paper, we consider a problem of learning to rank on a large-scal
e dataset with low-quality relevance labels aiming at maximizing the quality of
a trained ranker on a small validation dataset with high-quality ground truth re
levance labels. Motivated by the classical Gauss-Markov theorem for the linear r
egression problem, we formulate the problems of (1) reweighting training instanc
es and (2) remapping learning targets. We propose meta–gradient decision tree le
arning framework for optimizing weight and target functions by applying gradient
-based hyperparameter optimization. Experiments on a large-scale real-world data
set demonstrate that we can significantly improve state-of-the-art machine-learn
ing algorithms by incorporating our framework.
****

Discriminative Embeddings of Latent Variable Models for Structured Data
Hanjun Dai, Bo Dai, Le Song
Kernel classifiers and regressors designed for structured data, such as sequence
s, trees and graphs, have significantly advanced a number of interdisciplinary a
reas such as computational biology and drug design. Typically, kernels are desig
ned beforehand for a data type which either exploit statistics of the structures
 or make use of probabilistic generative models, and then a discriminative class
ifier is learned based on the kernels via convex optimization. However, such an
elegant two-stage approach also limited kernel methods from scaling up to millio
ns of data points, and exploiting discriminative information to learn feature re
presentations. We propose, structure2vec, an effective and scalable approach for
 structured data representation based on the idea of embedding latent variable m
odels into feature spaces, and learning such feature spaces using discriminative
 information. Interestingly, structure2vec extracts features by performing a seq
uence of function mappings in a way similar to graphical model inference procedu
res, such as mean field and belief propagation. In applications involving millio
ns of data points, we showed that structure2vec runs 2 times faster, produces mo
dels which are 10,000 times smaller, while at the same time achieving the state-
of-the-art predictive performance.
****

Robust Random Cut Forest Based Anomaly Detection on Streams
Sudipto Guha, Nina Mishra, Gourav Roy, Okke Schrijvers
In this paper we focus on the anomaly detection problem for dynamic data streams
 through the lens of random cut forests. We investigate a robust random cut data
 structure that can be used as a sketch or synopsis of the input stream. We prov
ide a plausible definition of non-parametric anomalies based on the influence of

an unseen point on the remainder of the data, i.e., the externality imposed by that point. We show how the sketch can be efficiently updated in a dynamic data stream. We demonstrate the viability of the algorithm on publicly available real data.

****

Training Neural Networks Without Gradients: A Scalable ADMM Approach

Gavin Taylor, Ryan Burmeister, Zheng Xu, Bharat Singh, Ankit Patel, Tom Goldstein

With the growing importance of large network models and enormous training datasets, GPUs have become increasingly necessary to train neural networks. This is largely because conventional optimization algorithms rely on stochastic gradient methods that don’t scale well to large numbers of cores in a cluster setting. Furthermore, the convergence of all gradient methods, including batch methods, suffers from common problems like saturation effects, poor conditioning, and saddle points. This paper explores an unconventional training method that uses alternating direction methods and Bregman iteration to train networks without gradient descent steps. The proposed method reduces the network training problem to a sequence of minimization sub-steps that can each be solved globally in closed form. The proposed method is advantageous because it avoids many of the caveats that make gradient methods slow on highly non-convex problems. In addition, the method exhibits strong scaling in the distributed setting, yielding linear speedups even when split over thousands of cores.

****

Clustering High Dimensional Categorical Data via Topographical Features

Chao Chen, Novi Quadrianto

Analysis of categorical data is a challenging task. In this paper, we propose to compute topographical features of high-dimensional categorical data. We propose an efficient algorithm to extract modes of the underlying distribution and their attractive basins. These topographical features provide a geometric view of the data and can be applied to visualization and clustering of real world challenging datasets. Experiments show that our principled method outperforms state-of-the-art clustering methods while also admits an embarrassingly parallel property.

****

Efficient Algorithms for Large-scale Generalized Eigenvector Computation and Canonical Correlation Analysis

Rong Ge, Chi Jin, Sham, Praneeth Netrapalli, Aaron Sidford

This paper considers the problem of canonical-correlation analysis (CCA) and, more broadly, the generalized eigenvector problem for a pair of symmetric matrices. These are two fundamental problems in data analysis and scientific computing with numerous applications in machine learning and statistics. We provide simple iterative algorithms, with improved runtimes, for solving these problems that are globally linearly convergent with moderate dependencies on the condition numbers and eigenvalue gaps of the matrices involved. We obtain our results by reducing CCA to the top-k generalized eigenvector problem. We solve this problem through a general framework that simply requires black box access to an approximate linear system solver. Instantiating this framework with accelerated gradient descent we obtain a running time of \order\fracz k \sqrtκρ \log(1/ε) \log \left(kκ/ρ \right) where z is the total number of nonzero entries, κis the condition number and ρis the relative eigenvalue gap of the appropriate matrices. Our algorithm is linear in the input size and the number of components k up to a \log(k) factor. This is essential for handling large-scale matrices that appear in practice. To the best of our knowledge this is the first such algorithm with global linear convergence. We hope that our results prompt further research and ultimately improve the practical running time for performing these important data analysis procedures on large data sets.

****

Algorithms for Optimizing the Ratio of Submodular Functions

Wenruo Bai, Rishabh Iyer, Kai Wei, Jeff Bilmes

We investigate a new optimization problem involving minimizing the Ratio of Submodular (RS) functions. We argue that this problem occurs naturally in several re

al world applications. We then show the connection between this problem and seve
ral related problems, including minimizing the difference of submodular function
s, and to submodular optimization subject to submodular constraints. We show RS
that optimization can be solved within bounded approximation factors. We also pr
ovide a hardness bound and show that our tightest algorithm matches the lower bo
und up to a \log factor. Finally, we empirically demonstrate the performance and
 good scalability properties of our algorithms.
****

Model-Free Imitation Learning with Policy Optimization
Jonathan Ho, Jayesh Gupta, Stefano Ermon
In imitation learning, an agent learns how to behave in an environment with an u
nknown cost function by mimicking expert demonstrations. Existing imitation lear
ning algorithms typically involve solving a sequence of planning or reinforcemen
t learning problems. Such algorithms are therefore not directly applicable to la
rge, high-dimensional environments, and their performance can significantly degr
ade if the planning problems are not solved to optimality. Under the apprentices
hip learning formalism, we develop alternative model-free algorithms for finding
 a parameterized stochastic policy that performs at least as well as an expert p
olicy on an unknown cost function, based on sample trajectories from the expert.
 Our approach, based on policy gradients, scales to large continuous environment
s with guaranteed convergence to local minima.
****

ADIOS: Architectures Deep In Output Space
Moustapha Cisse, Maruan Al-Shedivat, Samy Bengio
Multi-label classification is a generalization of binary classification where th
e task consists in predicting \emphsets of labels. With the availability of ever
 larger datasets, the multi-label setting has become a natural one in many appli
cations, and the interest in solving multi-label problems has grown significantl
y. As expected, deep learning approaches are now yielding state-of-the-art perfo
rmance for this class of problems. Unfortunately, they usually do not take into
account the often unknown but nevertheless rich relationships between labels. In
 this paper, we propose to make use of this underlying structure by learning to
partition the labels into a Markov Blanket Chain and then applying a novel deep
architecture that exploits the partition. Experiments on several popular and lar
ge multi-label datasets demonstrate that our approach not only yields significan
t improvements, but also helps to overcome trade-offs specific to the multi-labe
l classification setting.
****

Conditional Dependence via Shannon Capacity: Axioms, Estimators and Applications
Weihao Gao, Sreeram Kannan, Sewoong Oh, Pramod Viswanath
We consider axiomatically the problem of estimating the strength of a conditiona
l dependence relationship P_Y|X from a random variables X to a random variable Y
. This has applications in determining the strength of a known causal relationsh
ip, where the strength depends only on the conditional distribution of the effec
t given the cause (and not on the driving distribution of the cause). Shannon ca
pacity, appropriately regularized, emerges as a natural measure under these axio
ms. We examine the problem of calculating Shannon capacity from the observed sam
ples and propose a novel fixed-k nearest neighbor estimator, and demonstrate its
 consistency. Finally, we demonstrate an application to single-cell flow-cytomet
ry, where the proposed estimators significantly reduce sample complexity.
****

Control of Memory, Active Perception, and Action in Minecraft
Junhyuk Oh, Valliappa Chockalingam, Satinder, Honglak Lee
In this paper, we introduce a new set of reinforcement learning (RL) tasks in Mi
necraft (a flexible 3D world). We then use these tasks to systematically compare
 and contrast existing deep reinforcement learning (DRL) architectures with our
new memory-based DRL architectures. These tasks are designed to emphasize, in a
controllable manner, issues that pose challenges for RL methods including partia
l observability (due to first-person visual observations), delayed rewards, high
-dimensional visual observations, and the need to use active perception in a cor

rect manner so as to perform well in the tasks. While these tasks are conceptual
ly simple to describe, by virtue of having all of these challenges simultaneousl
y they are difficult for current DRL architectures. Additionally, we evaluate th
e generalization performance of the architectures on environments not used durin
g training. The experimental results show that our new architectures generalize
to unseen environments better than existing DRL architectures.
****

The Label Complexity of Mixed-Initiative Classifier Training
Jina Suh, Xiaojin Zhu, Saleema Amershi
Mixed-initiative classifier training, where the human teacher can choose which i
tems to label or to label items chosen by the computer, has enjoyed empirical su
ccess but without a rigorous statistical learning theoretical justification. We
analyze the label complexity of a simple mixed-initiative training mechanism usi
ng teach- ing dimension and active learning. We show that mixed-initiative train
ing is advantageous com- pared to either computer-initiated (represented by acti
ve learning) or human-initiated classifier training. The advantage exists across
 all human teaching abilities, from optimal to completely unhelpful teachers. We
 further improve classifier training by educating the human teachers. This is do
ne by showing, or explaining, optimal teaching sets to the human teachers. We co
nduct Mechanical Turk human experiments on two stylistic classifier training tas
ks to illustrate our approach.
****

Bayesian Poisson Tucker Decomposition for Learning the Structure of Internationa
l Relations
Aaron Schein, Mingyuan Zhou, David Blei, Hanna Wallach
We introduce Bayesian Poisson Tucker decomposition (BPTD) for modeling country–c
ountry interaction event data. These data consist of interaction events of the f
orm “country i took action a toward country j at time t.” BPTD discovers overlap
ping country–community memberships, including the number of latent communities.
In addition, it discovers directed community–community interaction networks that
 are specific to “topics” of action types and temporal “regimes.” We show that B
PTD yields an efficient MCMC inference algorithm and achieves better predictive
performance than related models. We also demonstrate that it discovers interpret
able latent structure that agrees with our knowledge of international relations.
****

Tensor Decomposition via Joint Matrix Schur Decomposition
Nicolo Colombo, Nikos Vlassis
We describe an approach to tensor decomposition that involves extracting a set o
f observable matrices from the tensor and applying an approximate joint Schur de
composition on those matrices, and we establish the corresponding first-order pe
rturbation bounds. We develop a novel iterative Gauss-Newton algorithm for joint
 matrix Schur decomposition, which minimizes a nonconvex objective over the mani
fold of orthogonal matrices, and which is guaranteed to converge to a global opt
imum under certain conditions. We empirically demonstrate that our algorithm is
faster and at least as accurate and robust than state-of-the-art algorithms for
this problem.
****

Continuous Deep Q-Learning with Model-based Acceleration
Shixiang Gu, Timothy Lillicrap, Ilya Sutskever, Sergey Levine
Model-free reinforcement learning has been successfully applied to a range of ch
allenging problems, and has recently been extended to handle large neural networ
k policies and value functions. However, the sample complexity of model-free alg
orithms, particularly when using high-dimensional function approximators, tends
to limit their applicability to physical systems. In this paper, we explore algo
rithms and representations to reduce the sample complexity of deep reinforcement
 learning for continuous control tasks. We propose two complementary techniques
for improving the efficiency of such algorithms. First, we derive a continuous v
ariant of the Q-learning algorithm, which we call normalized advantage functions
 (NAF), as an alternative to the more commonly used policy gradient and actor-cr
itic methods. NAF representation allows us to apply Q-learning with experience r

eplay to continuous tasks, and substantially improves performance on a set of si mulated robotic control tasks. To further improve the efficiency of our approach , we explore the use of learned models for accelerating model-free reinforcement learning. We show that iteratively refitted local linear models are especially effective for this, and demonstrate substantially faster learning on domains whe re such models are applicable.
****

Domain Adaptation with Conditional Transferable Components
Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, Bernhard Sc hölkopf
Domain adaptation arises in supervised learning when the training (source domain) and test (target domain) data have different distributions. Let X and Y denote the features and target, respectively, previous work on domain adaptation consi ders the covariate shift situation where the distribution of the features P(X) c hanges across domains while the conditional distribution P(Y|X) stays the same. To reduce domain discrepancy, recent methods try to find invariant components \m athcalT(X) that have similar P(\mathcalT(X)) by explicitly minimizing a distribu tion discrepancy measure. However, it is not clear if P(Y|\mathcalT(X)) in diffe rent domains is also similar when P(Y|X) changes. Furthermore, transferable comp onents do not necessarily have to be invariant. If the change in some components is identifiable, we can make use of such components for prediction in the targe t domain. In this paper, we focus on the case where P(X|Y) and P(Y) both change in a causal system in which Y is the cause for X. Under appropriate assumptions, we aim to extract conditional transferable components whose conditional distrib ution P(\mathcalT(X)|Y) is invariant after proper location-scale (LS) transforma tions, and identify how P(Y) changes between domains simultaneously. We provide theoretical analysis and empirical evaluation on both synthetic and real-world d ata to show the effectiveness of our method.
****

Fixed Point Quantization of Deep Convolutional Networks
Darryl Lin, Sachin Talathi, Sreekanth Annapureddy
In recent years increasingly complex architectures for deep convolution networks (DCNs) have been proposed to boost the performance on image recognition tasks. However, the gains in performance have come at a cost of substantial increase in computation and model storage resources. Fixed point implementation of DCNs has the potential to alleviate some of these complexities and facilitate potential deployment on embedded hardware. In this paper, we propose a quantizer design fo r fixed point implementation of DCNs. We formulate and solve an optimization pro blem to identify optimal fixed point bit-width allocation across DCN layers. Our experiments show that in comparison to equal bit-width settings, the fixed poin t DCNs with optimized bit width allocation offer >20% reduction in the model siz e without any loss in accuracy on CIFAR-10 benchmark. We also demonstrate that f ine-tuning can further enhance the accuracy of fixed point DCNs beyond that of t he original floating point model. In doing so, we report a new state-of-the-art fixed point performance of 6.78% error-rate on CIFAR-10 benchmark.
****

Provable Algorithms for Inference in Topic Models
Sanjeev Arora, Rong Ge, Frederic Koehler, Tengyu Ma, Ankur Moitra
Recently, there has been considerable progress on designing algorithms with prov able guarantees —typically using linear algebraic methods—for parameter learning in latent variable models. Designing provable algorithms for inference has prov ed more difficult. Here we take a first step towards provable inference in topic models. We leverage a property of topic models that enables us to construct sim ple linear estimators for the unknown topic proportions that have small variance , and consequently can work with short documents. Our estimators also correspond to finding an estimate around which the posterior is well-concentrated. We show lower bounds that for shorter documents it can be information theoretically imp ossible to find the hidden topics. Finally, we give empirical results that demon strate that our algorithm works on realistic topic models. It yields good soluti ons on synthetic data and runs in time comparable to a single iteration of Gibbs

sampling.
****
Epigraph projections for fast general convex programming
Po-Wei Wang, Matt Wytock, Zico Kolter
This paper develops an approach for efficiently solving general convex optimizat ion problems specified as disciplined convex programs (DCP), a common general-pu rpose modeling framework. Specifically we develop an algorithm based upon fast e pigraph projections, projections onto the epigraph of a convex function, an appr oach closely linked to proximal operator methods. We show that by using these op erators, we can solve any disciplined convex program without transforming the pr oblem to a standard cone form, as is done by current DCP libraries. We then deve lop a large library of efficient epigraph projection operators, mirroring and ex tending work on fast proximal algorithms, for many common convex functions. Fina lly, we evaluate the performance of the algorithm, and show it often achieves or der of magnitude speedups over existing general-purpose optimization solvers.
****
Fast Algorithms for Segmented Regression
Jayadev Acharya, Ilias Diakonikolas, Jerry Li, Ludwig Schmidt
We study the fixed design segmented regression problem: Given noisy samples from a piecewise linear function f, we want to recover f up to a desired accuracy in mean-squared error. Previous rigorous approaches for this problem rely on dynam ic programming (DP) and, while sample efficient, have running time quadratic in the sample size. As our main contribution, we provide new sample near-linear tim e algorithms for the problem that - while not being minimax optimal - achieve a significantly better sample-time tradeoff on large datasets compared to the DP a pproach. Our experimental evaluation shows that, compared with the DP approach, our algorithms provide a convergence rate that is only off by a factor of 2 to 4 , while achieving speedups of three orders of magnitude.
****
Energetic Natural Gradient Descent
Philip Thomas, Bruno Castro Silva, Christoph Dann, Emma Brunskill
We propose a new class of algorithms for minimizing or maximizing functions of p arametric probabilistic models. These new algorithms are natural gradient algori thms that leverage more information than prior methods by using a new metric ten sor in place of the commonly used Fisher information matrix. This new metric ten sor is derived by computing directions of steepest ascent where the distance bet ween distributions is measured using an approximation of energy distance (as opp osed to Kullback-Leibler divergence, which produces the Fisher information matri x), and so we refer to our new ascent direction as the energetic natural gradien t.
****
Partition Functions from Rao-Blackwellized Tempered Sampling
David Carlson, Patrick Stinson, Ari Pakman, Liam Paninski
Partition functions of probability distributions are important quantities for mo del evaluation and comparisons. We present a new method to compute partition fun ctions of complex and multimodal distributions. Such distributions are often sam pled using simulated tempering, which augments the target space with an auxiliar y inverse temperature variable. Our method exploits the multinomial probability law of the inverse temperatures, and provides estimates of the partition functio n in terms of a simple quotient of Rao-Blackwellized marginal inverse temperatur e probability estimates, which are updated while sampling. We show that the meth od has interesting connections with several alternative popular methods, and off ers some significant advantages. In particular, we empirically find that the new method provides more accurate estimates than Annealed Importance Sampling when calculating partition functions of large Restricted Boltzmann Machines (RBM); mo reover, the method is sufficiently accurate to track training and validation log -likelihoods during learning of RBMs, at minimal computational cost.
****
Learning Mixtures of Plackett-Luce Models
Zhibing Zhao, Peter Piech, Lirong Xia

In this paper we address the identifiability and efficient learning problems of finite mixtures of Plackett-Luce models for rank data. We prove that for any k≥2 , the mixture of k Plackett-Luce models for no more than 2k-1 alternatives is no n-identifiable and this bound is tight for k=2. For generic identifiability, we prove that the mixture of k Plackett-Luce models over m alternatives is \em gene rically identifiable if k≤■\frac m-2 2■!. We also propose an efficient generaliz ed method of moments (GMM) algorithm to learn the mixture of two Plackett-Luce m odels and show that the algorithm is consistent. Our experiments show that our G MM algorithm is significantly faster than the EMM algorithm by Gormley & Murphy (2008), while achieving competitive statistical efficiency.
****

Near Optimal Behavior via Approximate State Abstraction
David Abel, David Hershkowitz, Michael Littman
The combinatorial explosion that plagues planning and reinforcement learning (RL) algorithms can be moderated using state abstraction. Prohibitively large task representations can be condensed such that essential information is preserved, a nd consequently, solutions are tractably computable. However, exact abstractions , which treat only fully-identical situations as equivalent, fail to present opp ortunities for abstraction in environments where no two situations are exactly a like. In this work, we investigate approximate state abstractions, which treat n early-identical situations as equivalent. We present theoretical guarantees of t he quality of behaviors derived from four types of approximate abstractions. Add itionally, we empirically demonstrate that approximate abstractions lead to redu ction in task complexity and bounded loss of optimality of behavior in a variety of environments.
****

Power of Ordered Hypothesis Testing
Lihua Lei, William Fithian
Ordered testing procedures are multiple testing procedures that exploit a pre-sp ecified ordering of the null hypotheses, from most to least promising. We analyz e and compare the power of several recent proposals using the asymptotic framewo rk of Li & Barber (2015). While accumulation tests including ForwardStop can be quite powerful when the ordering is very informative, they are asymptotically po werless when the ordering is weaker. By contrast, Selective SeqStep, proposed by Barber & Candes (2015), is much less sensitive to the quality of the ordering. We compare the power of these procedures in different regimes, concluding that S elective SeqStep dominates accumulation tests if either the ordering is weak or non-null hypotheses are sparse or weak. Motivated by our asymptotic analysis, we derive an improved version of Selective SeqStep which we call Adaptive SeqStep, analogous to Storey’s improvement on the Benjamini-Hochberg procedure. We compa re these methods using the GEO-Query data set analyzed by (Li & Barber, 2015) an d find Adaptive SeqStep has favorable performance for both good and bad prior or derings.
****

PHOG: Probabilistic Model for Code
Pavol Bielik, Veselin Raychev, Martin Vechev
We introduce a new generative model for code called probabilistic higher order g rammar (PHOG). PHOG generalizes probabilistic context free grammars (PCFGs) by a llowing conditioning of a production rule beyond the parent non-terminal, thus c apturing rich contexts relevant to programs. Even though PHOG is more powerful t han a PCFG, it can be learned from data just as efficiently. We trained a PHOG m odel on a large JavaScript code corpus and show that it is more precise than exi sting models, while similarly fast. As a result, PHOG can immediately benefit ex isting programming tools based on probabilistic models of code.
****

Shifting Regret, Mirror Descent, and Matrices
Andras Gyorgy, Csaba Szepesvari
We consider the problem of online prediction in changing environments. In this f ramework the performance of a predictor is evaluated as the loss relative to an arbitrarily changing predictor, whose individual components come from a base cla

ss of predictors. Typical results in the literature consider different base clas ses (experts, linear predictors on the simplex, etc.) separately. Introducing an arbitrary mapping inside the mirror decent algorithm, we provide a framework th at unifies and extends existing results. As an example, we prove new shifting re gret bounds for matrix prediction problems.

****

Scalable Gradient-Based Tuning of Continuous Regularization Hyperparameters

Jelena Luketina, Mathias Berglund, Klaus Greff, Tapani Raiko

Hyperparameter selection generally relies on running multiple full training tria ls, with selection based on validation set performance. We propose a gradient-ba sed approach for locally adjusting hyperparameters during training of the model. Hyperparameters are adjusted so as to make the model parameter gradients, and h ence updates, more advantageous for the validation cost. We explore the approach for tuning regularization hyperparameters and find that in experiments on MNIST , SVHN and CIFAR-10, the resulting regularization levels are within the optimal regions. The additional computational cost depends on how frequently the hyperpa rameters are trained, but the tested scheme adds only 30% computational overhead regardless of the model size. Since the method is significantly less computatio nally demanding compared to similar gradient-based approaches to hyperparameter optimization, and consistently finds good hyperparameter values, it can be a use ful tool for training neural network models.

****

Model-Free Trajectory Optimization for Reinforcement Learning

Riad Akrour, Gerhard Neumann, Hany Abdulsamad, Abbas Abdolmaleki

Many of the recent Trajectory Optimization algorithms alternate between local ap proximation of the dynamics and conservative policy update. However, linearly ap proximating the dynamics in order to derive the new policy can bias the update a nd prevent convergence to the optimal policy. In this article, we propose a new model-free algorithm that backpropagates a local quadratic time-dependent Q-Func tion, allowing the derivation of the policy update in closed form. Our policy up date ensures exact KL-constraint satisfaction without simplifying assumptions on the system dynamics demonstrating improved performance in comparison to related Trajectory Optimization algorithms linearizing the dynamics.

****

Controlling the distance to a Kemeny consensus without computing it

Yunlong Jiao, Anna Korba, Eric Sibony

Due to its numerous applications, rank aggregation has become a problem of major interest across many fields of the computer science literature. In the vast maj ority of situations, Kemeny consensus(es) are considered as the ideal solutions. It is however well known that their computation is NP-hard. Many contributions have thus established various results to apprehend this complexity. In this pape r we introduce a practical method to predict, for a ranking and a dataset, how c lose the Kemeny consensus(es) are to this ranking. A major strength of this meth od is its generality: it does not require any assumption on the dataset nor the ranking. Furthermore, it relies on a new geometric interpretation of Kemeny aggr egation that, we believe, could lead to many other results.

****

Horizontally Scalable Submodular Maximization

Mario Lucic, Olivier Bachem, Morteza Zadimoghaddam, Andreas Krause

A variety of large-scale machine learning problems can be cast as instances of c onstrained submodular maximization. Existing approaches for distributed submodul ar maximization have a critical drawback: The capacity - number of instances tha t can fit in memory - must grow with the data set size. In practice, while one c an provision many machines, the capacity of each machine is limited by physical constraints. We propose a truly scalable approach for distributed submodular max imization under fixed capacity. The proposed framework applies to a broad class of algorithms and constraints and provides theoretical guarantees on the approxi mation factor for any available capacity. We empirically evaluate the proposed a lgorithm on a variety of data sets and demonstrate that it achieves performance competitive with the centralized greedy solution.

****
Group Equivariant Convolutional Networks
Taco Cohen, Max Welling
We introduce Group equivariant Convolutional Neural Networks (G-CNNs), a natural generalization of convolutional neural networks that reduces sample complexity by exploiting symmetries. G-CNNs use G-convolutions, a new type of layer that enjoys a substantially higher degree of weight sharing than regular convolution layers. G-convolutions increase the expressive capacity of the network without increasing the number of parameters. Group convolution layers are easy to use and can be implemented with negligible computational overhead for discrete groups generated by translations, reflections and rotations. G-CNNs achieve state of the art results on CIFAR10 and rotated MNIST.
****
Stochastic Discrete Clenshaw-Curtis Quadrature
Nico Piatkowski, Katharina Morik
The partition function is fundamental for probabilistic graphical models—it is required for inference, parameter estimation, and model selection. Evaluating this function corresponds to discrete integration, namely a weighted sum over an exponentially large set. This task quickly becomes intractable as the dimensionality of the problem increases. We propose an approximation scheme that, for any discrete graphical model whose parameter vector has bounded norm, estimates the partition function with arbitrarily small error. Our algorithm relies on a near minimax optimal polynomial approximation to the potential function and a Clenshaw-Curtis style quadrature. Furthermore, we show that this algorithm can be randomized to split the computation into a high-complexity part and a low-complexity part, where the latter may be carried out on small computational devices. Experiments confirm that the new randomized algorithm is highly accurate if the parameter norm is small, and is otherwise comparable to methods with unbounded error.
****
Correcting Forecasts with Multifactor Neural Attention
Matthew Riemer, Aditya Vempaty, Flavio Calmon, Fenno Heath, Richard Hull, Elham Khabiri
Automatic forecasting of time series data is a challenging problem in many industries. Current forecast models adopted by businesses do not provide adequate means for including data representing external factors that may have a significant impact on the time series, such as weather, national events, local events, social media trends, promotions, etc. This paper introduces a novel neural network attention mechanism that naturally incorporates data from multiple external sources without the feature engineering needed to get other techniques to work. We demonstrate empirically that the proposed model achieves superior performance for predicting the demand of 20 commodities across 107 stores of one of America’s largest retailers when compared to other baseline models, including neural networks, linear models, certain kernel methods, Bayesian regression, and decision trees. Our method ultimately accounts for a 23.9% relative improvement as a result of the incorporation of external data sources, and provides an unprecedented level of descriptive ability for a neural network forecasting model.
****
Learning Representations for Counterfactual Inference
Fredrik Johansson, Uri Shalit, David Sontag
Observational studies are rising in importance due to the widespread accumulation of data in fields such as healthcare, education, employment and ecology. We consider the task of answering counterfactual questions such as, “Would this patient have lower blood sugar had she received a different medication?". We propose a new algorithmic framework for counterfactual inference which brings together ideas from domain adaptation and representation learning. In addition to a theoretical justification, we perform an empirical comparison with previous approaches to causal inference from observational data. Our deep learning algorithm significantly outperforms the previous state-of-the-art.
****
Automatic Construction of Nonparametric Relational Regression Models for Multipl

e Time Series
Yunseong Hwang, Anh Tong, Jaesik Choi
Gaussian Processes (GPs) provide a general and analytically tractable way of mod eling complex time-varying, nonparametric functions. The Automatic Bayesian Cova riance Discovery (ABCD) system constructs natural-language description of time-s eries data by treating unknown time-series data nonparametrically using GP with a composite covariance kernel function. Unfortunately, learning a composite cova riance kernel with a single time-series data set often results in less informati ve kernel that may not give qualitative, distinctive descriptions of data. We ad dress this challenge by proposing two relational kernel learning methods which c an model multiple time-series data sets by finding common, shared causes of chan ges. We show that the relational kernel learning methods find more accurate mode ls for regression problems on several real-world data sets; US stock data, US ho use price index data and currency exchange rate data.
****
Inference Networks for Sequential Monte Carlo in Graphical Models
Brooks Paige, Frank Wood
We introduce a new approach for amortizing inference in directed graphical model s by learning heuristic approximations to stochastic inverses, designed specific ally for use as proposal distributions in sequential Monte Carlo methods. We des cribe a procedure for constructing and learning a structured neural network whic h represents an inverse factorization of the graphical model, resulting in a con ditional density estimator that takes as input particular values of the observed random variables, and returns an approximation to the distribution of the laten t variables. This recognition model can be learned offline, independent from any particular dataset, prior to performing inference. The output of these networks can be used as automatically-learned high-quality proposal distributions to acc elerate sequential Monte Carlo across a diverse range of problem settings.
****
Slice Sampling on Hamiltonian Trajectories
Benjamin Bloem-Reddy, John Cunningham
Hamiltonian Monte Carlo and slice sampling are amongst the most widely used and studied classes of Markov Chain Monte Carlo samplers. We connect these two metho ds and present Hamiltonian slice sampling, which allows slice sampling to be car ried out along Hamiltonian trajectories, or transformations thereof. Hamiltonian slice sampling clarifies a class of model priors that induce closed-form slice samplers. More pragmatically, inheriting properties of slice samplers, it offers advantages over Hamiltonian Monte Carlo, in that it has fewer tunable hyperpara meters and does not require gradient information. We demonstrate the utility of Hamiltonian slice sampling out of the box on problems ranging from Gaussian proc ess regression to Pitman-Yor based mixture models.
****
Noisy Activation Functions
Caglar Gulcehre, Marcin Moczulski, Misha Denil, Yoshua Bengio
Common nonlinear activation functions used in neural networks can cause training difficulties due to the saturation behavior of the activation function, which m ay hide dependencies that are not visible to vanilla-SGD (using first order grad ients only). Gating mechanisms that use softly saturating activation functions t o emulate the discrete switching of digital logic circuits are good examples of this. We propose to exploit the injection of appropriate noise so that the gradi ents may flow easily, even if the noiseless application of the activation functi on would yield zero gradients. Large noise will dominate the noise-free gradient and allow stochastic gradient descent to explore more. By adding noise only to the problematic parts of the activation function, we allow the optimization proc edure to explore the boundary between the degenerate saturating) and the well-be haved parts of the activation function. We also establish connections to simulat ed annealing, when the amount of noise is annealed down, making it easier to opt imize hard objective functions. We find experimentally that replacing such satur ating activation functions by noisy variants helps optimization in many contexts , yielding state-of-the-art or competitive results on different datasets and tas

k, especially when training seems to be the most difficult, e.g., when curriculu
m learning is necessary to obtain good results.
****
PD-Sparse : A Primal and Dual Sparse Approach to Extreme Multiclass and Multilab
el Classification
Ian En-Hsu Yen, Xiangru Huang, Pradeep Ravikumar, Kai Zhong, Inderjit Dhillon
We consider Multiclass and Multilabel classification with extremely large number
 of classes, of which only few are labeled to each instance. In such setting, st
andard methods that have training, prediction cost linear to the number of class
es become intractable. State-of-the-art methods thus aim to reduce the complexit
y by exploiting correlation between labels under assumption that the similarity
between labels can be captured by structures such as low-rank matrix or balanced
 tree. However, as the diversity of labels increases in the feature space, struc
tural assumption can be easily violated, which leads to degrade in the testing p
erformance. In this work, we show that a margin-maximizing loss with l1 penalty,
 in case of Extreme Classification, yields extremely sparse solution both in pri
mal and in dual without sacrificing the expressive power of predictor. We thus p
ropose a Fully-Corrective Block-Coordinate Frank-Wolfe (FC-BCFW) algorithm that
exploits both primal and dual sparsity to achieve a complexity sublinear to the
number of primal and dual variables. A bi-stochastic search method is proposed t
o further improve the efficiency. In our experiments on both Multiclass and Mult
ilabel problems, the proposed method achieves significant higher accuracy than e
xisting approaches of Extreme Classification with very competitive training and
prediction time.
****