Quartz: Randomized Dual Coordinate Ascent with Arbitrary Sampling
Zheng Qu, Peter Richtarik, Tong Zhang

We study the problem of minimizing the average of a large number of smooth convex functions penalized with a strongly convex regularizer. We propose and analyze a novel primal-dual method (Quartz) which at every iteration samples and updates a random subset of the dual variables, chosen according to an arbitrary distribution. In contrast to typical analysis, we directly bound the decrease of the primal-dual error (in expectation), without the need to first analyze the dual error. Depending on the choice of the sampling, we obtain efficient serial and mini-batch variants of the method. In the serial case, our bounds match the best known bounds for SDCA (both with uniform and importance sampling). With standard mini-batching, our bounds predict initial data-independent speedup as well as additional data-driven speedup which depends on spectral and sparsity properties of the data.

****

Associative Memory via a Sparse Recovery Model
Arya Mazumdar, Ankit Singh Rawat

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Policy Gradient for Coherent Risk Measures
Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, Shie Mannor

Several authors have recently developed risk-sensitive policy gradient methods that augment the standard expected cost minimization problem with a measure of variability in cost. These studies have focused on specific risk-measures, such as the variance or conditional value at risk (CVaR). In this work, we extend the policy gradient method to the whole class of coherent risk measures, which is widely accepted in finance and operations research, among other fields. We consider both static and time-consistent dynamic risk measures. For static risk measures, our approach is in the spirit of policy gradient algorithms and combines a standard sampling approach with convex programming. For dynamic risk measures, our approach is actor-critic style and involves explicit approximation of value function. Most importantly, our contribution presents a unified approach to risk-sensitive reinforcement learning that generalizes and extends previous results.

****

A fast, universal algorithm to learn parametric nonlinear embeddings
Miguel A. Carreira-Perpinan, Max Vladymyrov

Nonlinear embedding algorithms such as stochastic neighbor embedding do dimensionality reduction by optimizing an objective function involving similarities between pairs of input patterns. The result is a low-dimensional projection of each input pattern. A common way to define an out-of-sample mapping is to optimize the objective directly over a parametric mapping of the inputs, such as a neural net. This can be done using the chain rule and a nonlinear optimizer, but is very slow, because the objective involves a quadratic number of terms each dependent on the entire mapping's parameters. Using the method of auxiliary coordinates, we derive a training algorithm that works by alternating steps that train an auxiliary embedding with steps that train the mapping. This has two advantages: 1) The algorithm is universal in that a specific learning algorithm for any choice of embedding and mapping can be constructed by simply reusing existing algorithms for the embedding and for the mapping. A user can then try possible mappings and embeddings with less effort. 2) The algorithm is fast, and it can reuse N-body methods developed for nonlinear embeddings, yielding linear-time iterations.

****

Stochastic Online Greedy Learning with Semi-bandit Feedbacks
Tian Lin, Jian Li, Wei Chen

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth

ors prior to requesting a name change in the electronic proceedings.
****

SubmodBoxes: Near-Optimal Search for a Set of Diverse Object Proposals
Qing Sun, Dhruv Batra

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Robust Portfolio Optimization
Huitong Qiu, Fang Han, Han Liu, Brian Caffo

We propose a robust portfolio optimization approach based on quantile statistics. The proposed method is robust to extreme events in asset returns, and accommodates large portfolios under limited historical data. Specifically, we show that the risk of the estimated portfolio converges to the oracle optimal risk with parametric rate under weakly dependent asset returns. The theory does not rely on higher order moment assumptions, thus allowing for heavy-tailed asset returns. Moreover, the rate of convergence quantifies that the size of the portfolio under management is allowed to scale exponentially with the sample size of the historical data. The empirical effectiveness of the proposed method is demonstrated under both synthetic and real stock data. Our work extends existing ones by achieving robustness in high dimensions, and by allowing serial dependence.
****

Top-k Multiclass SVM
Maksim Lapin, Matthias Hein, Bernt Schiele

Class ambiguity is typical in image classification problems with a large number of classes. When classes are difficult to discriminate, it makes sense to allow k guesses and evaluate classifiers based on the top-k error instead of the standard zero-one loss. We propose top-k multiclass SVM as a direct method to optimize for top-k performance. Our generalization of the well-known multiclass SVM is based on a tight convex upper bound of the top-k error. We propose a fast optimization scheme based on an efficient projection onto the top-k simplex, which is of its own interest. Experiments on five datasets show consistent improvements in top-k accuracy compared to various baselines.
****

Less is More: Nyström Computational Regularization
Alessandro Rudi, Raffaello Camoriano, Lorenzo Rosasco

We study Nyström type subsampling approaches to large   scale kernel methods, and prove   learning bounds in the statistical learning setting, where random sampling and high probability estimates are considered.   In particular, we prove that these approaches can achieve optimal learning bounds, provided the subsampling level is suitably chosen. These results suggest a simple incremental variant of Nyström kernel ridge regression, where the subsampling level controls at the same time regularization and computations. Extensive experimental analysis shows that the considered approach achieves state of the art performances on benchmark large scale datasets.
****

Parallel Recursive Best-First AND/OR Search for Exact MAP Inference in Graphical Models
Akihiro Kishimoto, Radu Marinescu, Adi Botea

The paper presents and evaluates the power of parallel search for exact MAP inference in graphical models. We introduce a new parallel shared-memory recursive best-first AND/OR search algorithm, called SPRBFAOO, that explores the search space in a best-first manner while operating with restricted memory. Our experiments show that SPRBFAOO is often superior to the current state-of-the-art sequential AND/OR search approaches, leading to considerable speed-ups (up to 7-fold with 12 threads), especially on hard problem instances.
****

Differentially private subspace clustering
Yining Wang, Yu-Xiang Wang, Aarti Singh

Subspace clustering is an unsupervised learning problem that aims at grouping data points into multiple clusters'' so that data points in a single cluster lie approximately on a low-dimensional linear subspace. It is originally motivated by 3D motion segmentation in computer vision, but has recently been generically applied to a wide range of statistical machine learning problems, which often involves sensitive datasets about human subjects. This raises a dire concern for data privacy. In this work, we build on the framework ofdifferential privacy'' and present two provably private subspace clustering algorithms. We demonstrate via both theory and experiments that one of the presented methods enjoys formal privacy and utility guarantees; the other one asymptotically preserves differential privacy while having good performance in practice. Along the course of the proof, we also obtain two new provable guarantees for the agnostic subspace clustering and the graph connectivity problem which might be of independent interests.

****

Matrix Completion with Noisy Side Information

Kai-Yang Chiang, Cho-Jui Hsieh, Inderjit S. Dhillon

We study matrix completion problem with side information. Side information has been considered in several matrix completion applications, and is generally shown to be useful empirically. Recently, Xu et al. studied the effect of side information for matrix completion under a theoretical viewpoint, showing that sample complexity can be significantly reduced given completely clean features. However, since in reality most given features are noisy or even weakly informative, how to develop a general model to handle general feature set, and how much the noisy features can help matrix recovery in theory, is still an important issue to investigate. In this paper, we propose a novel model that balances between features and observations simultaneously, enabling us to leverage feature information yet to be robust to feature noise. Moreover, we study the effectof general features in theory, and show that by using our model, the sample complexity can still be lower than matrix completion as long as features are sufficiently informative. This result provides a theoretical insight of usefulness for general side information. Finally, we consider synthetic data and two real applications - relationship prediction and semi-supervised clustering, showing that our model outperforms other methods for matrix completion with features both in theory and practice.

****

Nonparametric von Mises Estimators for Entropies, Divergences and Mutual Informations

Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, james m. robins

We propose and analyse estimators for statistical functionals of one or moredistributions under nonparametric assumptions.Our estimators are derived from the von Mises expansion andare based on the theory of influence functions, which appearin the semiparametric statistics literature.We show that estimators based either on data-splitting or a leave-one-out techniqueenjoy fast rates of convergence and other favorable theoretical properties.We apply this framework to derive estimators for several popular informationtheoretic quantities, and via empirical evaluation, show the advantage of thisapproach over existing estimators.

****

Semi-Supervised Factored Logistic Regression for High-Dimensional Neuroimaging Data

Danilo Bzdok, Michael Eickenberg, Olivier Grisel, Bertrand Thirion, Gael Varoquaux

Imaging neuroscience links human behavior to aspects of brain biology in ever-increasing datasets. Existing neuroimaging methods typically perform either discovery of unknown neural structure or testing of neural structure associated with mental tasks. However, testing hypotheses on the neural correlates underlying larger sets of mental tasks necessitates adequate representations for the observations. We therefore propose to blend representation modelling and task classification into a unified statistical learning problem. A multinomial logistic regression is introduced that is constrained by factored coefficients and coupled with a n autoencoder. We show that this approach yields more accurate and interpretable neural models of psychological tasks in a reference dataset, as well as better generalization to other datasets.

****

Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting

Xingjian SHI, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, Wang-chun WOO

The goal of precipitation nowcasting is to predict the future rainfall intensity in a local region over a relatively short period of time. Very few previous studies have examined this crucial and challenging weather forecasting problem from the machine learning perspective. In this paper, we formulate precipitation nowcasting as a spatiotemporal sequence forecasting problem in which both the input and the prediction target are spatiotemporal sequences. By extending the fully connected LSTM (FC-LSTM) to have convolutional structures in both the input-to-state and state-to-state transitions, we propose the convolutional LSTM (ConvLSTM) and use it to build an end-to-end trainable model for the precipitation nowcasting problem. Experiments show that our ConvLSTM network captures spatiotemporal correlations better and consistently outperforms FC-LSTM and the state-of-the-art operational ROVER algorithm for precipitation nowcasting.

****

Infinite Factorial Dynamical Model

Isabel Valera, Francisco Ruiz, Lennart Svensson, Fernando Perez-Cruz

We propose the infinite factorial dynamic model (iFDM), a general Bayesian nonparametric model for source separation. Our model builds on the Markov Indian buffet process to consider a potentially unbounded number of hidden Markov chains (sources) that evolve independently according to some dynamics, in which the state space can be either discrete or continuous. For posterior inference, we develop an algorithm based on particle Gibbs with ancestor sampling that can be efficiently applied to a wide range of source separation problems. We evaluate the performance of our iFDM on four well-known applications: multitarget tracking, cocktail party, power disaggregation, and multiuser detection. Our experimental results show that our approach for source separation does not only outperform previous approaches, but it can also handle problems that were computationally intractable for existing approaches.

****

Dependent Multinomial Models Made Easy: Stick-Breaking with the Polya-gamma Augmentation

Scott Linderman, Matthew J. Johnson, Ryan P. Adams

Many practical modeling problems involve discrete data that are best represented as draws from multinomial or categorical distributions. For example, nucleotides in a DNA sequence, children's names in a given state and year, and text documents are all commonly modeled with multinomial distributions. In all of these cases, we expect some form of dependency between the draws: the nucleotide at one position in the DNA strand may depend on the preceding nucleotides, children's names are highly correlated from year to year, and topics in text may be correlated and dynamic. These dependencies are not naturally captured by the typical Dirichlet-multinomial formulation. Here, we leverage a logistic stick-breaking representation and recent innovations in P\'{o}lya-gamma augmentation to reformulate the multinomial distribution in terms of latent variables with jointly Gaussian likelihoods, enabling us to take advantage of a host of Bayesian inference techniques for Gaussian models with minimal overhead.

****

Sparse Linear Programming via Primal and Dual Augmented Coordinate Descent

Ian En-Hsu Yen, Kai Zhong, Cho-Jui Hsieh, Pradeep K. Ravikumar, Inderjit S. Dhillon

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Data Generation as Sequential Decision Making

Philip Bachman, Doina Precup

We connect a broad class of generative models through their shared reliance on sequential decision making. Motivated by this view, we develop extensions to an existing model, and then explore the idea further in the context of data imputation -- perhaps the simplest setting in which to investigate the relation between unconditional and conditional generative modelling. We formulate data imputation as an MDP and develop models capable of representing effective policies for it. We construct the models using neural networks and train them using a form of guided policy search. Our models generate predictions through an iterative process of feedback and refinement. We show that this approach can learn effective policies for imputation problems of varying difficulty and across multiple datasets.

****

Online Gradient Boosting

Alina Beygelzimer, Elad Hazan, Satyen Kale, Haipeng Luo

We extend the theory of boosting for regression problems to the online learning setting. Generalizing from the batch setting for boosting, the notion of a weak learning algorithm is modeled as an online learning algorithm with linear loss functions that competes with a base class of regression functions, while a strong learning algorithm is an online learning algorithm with smooth convex loss functions that competes with a larger class of regression functions. Our main result is an online gradient boosting algorithm which converts a weak online learning algorithm into a strong one where the larger class of functions is the linear span of the base class. We also give a simpler boosting algorithm that converts a weak online learning algorithm into a strong one where the larger class of functions is the convex hull of the base class, and prove its optimality.

****

Optimal Ridge Detection using Coverage Risk

Yen-Chi Chen, Christopher R. Genovese, Shirley Ho, Larry Wasserman

We introduce the concept of coverage risk as an error measure for density ridge estimation.The coverage risk generalizes the mean integrated square error to set estimation.We propose two risk estimators for the coverage risk and we show that we can select tuning parameters by minimizing the estimated risk.We study the rate of convergence for coverage risk and prove consistency of the risk estimators.We apply our method to three simulated datasets and to cosmology data.In all the examples, the proposed method successfully recover the underlying density structure.

****

A Tractable Approximation to Optimal Point Process Filtering: Application to Neural Encoding

Yuval Harel, Ron Meir, Manfred Opper

The process of dynamic state estimation (filtering) based on point process observations is in general intractable. Numerical sampling techniques are often practically useful, but lead to limited conceptual insight about optimal encoding/decoding strategies, which are of significant relevance to Computational Neuroscience. We develop an analytically tractable Bayesian approximation to optimal filtering based on point process observations, which allows us to introduce distributional assumptions about sensory cell properties, that greatly facilitates the analysis of optimal encoding in situations deviating from common assumptions of uniform coding. The analytic framework leads to insights which are difficult to obtain from numerical algorithms, and is consistent with experiments about the distribution of tuning curve centers. Interestingly, we find that the information gained from the absence of spikes may be crucial to performance.

****

Barrier Frank-Wolfe for Marginal Inference

Rahul G. Krishnan, Simon Lacoste-Julien, David Sontag

We introduce a globally-convergent algorithm for optimizing the tree-reweighted (TRW) variational objective over the marginal polytope. The algorithm is based on the conditional gradient method (Frank-Wolfe) and moves pseudomarginals within the marginal polytope through repeated maximum a posteriori (MAP) calls. This modular structure enables us to leverage black-box MAP solvers (both exact and approximate) for variational inference, and obtains more accurate results than tree-reweighted algorithms that optimize over the local consistency relaxation. Theoretically, we bound the sub-optimality for the proposed algorithm despite the TRW objective having unbounded gradients at the boundary of the marginal polytope. Empirically, we demonstrate the increased quality of results found by tightening the relaxation over the marginal polytope as well as the spanning tree polytope on synthetic and real-world instances.

****

Combinatorial Bandits Revisited

Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, marc lelarge

This paper investigates stochastic and adversarial combinatorial multi-armed bandit problems. In the stochastic setting under semi-bandit feedback, we derive a problem-specific regret lower bound, and discuss its scaling with the dimension of the decision space. We propose ESCB, an algorithm that efficiently exploits the structure of the problem and provide a finite-time analysis of its regret. ESCB has better performance guarantees than existing algorithms, and significantly outperforms these algorithms in practice. In the adversarial setting under bandit feedback, we propose CombEXP, an algorithm with the same regret scaling as state-of-the-art algorithms, but with lower computational complexity for some combinatorial problems.

****

Efficient and Parsimonious Agnostic Active Learning

Tzu-Kuo Huang, Alekh Agarwal, Daniel J. Hsu, John Langford, Robert E. Schapire

We develop a new active learning algorithm for the streaming settingsatisfying three important properties: 1) It provably works for anyclassifier representation and classification problem including thosewith severe noise. 2) It is efficiently implementable with an ERMoracle. 3) It is more aggressive than all previous approachessatisfying 1 and 2. To do this, we create an algorithm based on a newlydefined optimization problem and analyze it. We also conduct the firstexperimental analysis of all efficient agnostic active learningalgorithms, evaluating their strengths and weaknesses in differentsettings.

****

Policy Evaluation Using the Ω-Return

Philip S. Thomas, Scott Niekum, Georgios Theocharous, George Konidaris

We propose the Ω-return as an alternative to the λ-return currently used by the TD(λ) family of algorithms. The benefit of the Ω-return is that it accounts for the correlation of different length returns. Because it is difficult to compute exactly, we suggest one way of approximating the Ω-return. We provide empirical studies that suggest that it is superior to the λ-return and γ-return for a variety of problems.

****

Bayesian Optimization with Exponential Convergence

Kenji Kawaguchi, Leslie Pack Kaelbling, Tomás Lozano-Pérez

This paper presents a Bayesian optimization method with exponential convergence without the need of auxiliary optimization and without the delta-cover sampling. Most Bayesian optimization methods require auxiliary optimization: an additional non-convex global optimization problem, which can be time-consuming and hard to implement in practice. Also, the existing Bayesian optimization method with exponential convergence requires access to the delta-cover sampling, which was considered to be impractical. Our approach eliminates both requirements and achieves an exponential convergence rate.

****

Statistical Model Criticism using Kernel Two Sample Tests

James R. Lloyd, Zoubin Ghahramani

We propose an exploratory approach to statistical model criticism using maximum mean discrepancy (MMD) two sample tests. Typical approaches to model criticism require a practitioner to select a statistic by which to measure discrepancies be tween data and a statistical model. MMD two sample tests are instead constructed as an analytic maximisation over a large space of possible statistics and therefore automatically select the statistic which most shows any discrepancy. We demonstrate on synthetic data that the selected statistic, called the witness function, can be used to identify where a statistical model most misrepresents the data it was trained on. We then apply the procedure to real data where the models being assessed are restricted Boltzmann machines, deep belief networks and Gaussian process regression and demonstrate the ways in which these models fail to capture the properties of the data they are trained on.

****

Attention-Based Models for Speech Recognition

Jan K. Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, Yoshua Bengio

Recurrent sequence generators conditioned on input data through an attention mechanism have recently shown very good performance on a range of tasks including machine translation, handwriting synthesis and image caption generation. We extend the attention-mechanism with features needed for speech recognition. We show that while an adaptation of the model used for machine translation reaches a competitive 18.6\% phoneme error rate (PER) on the TIMIT phoneme recognition task, it can only be applied to utterances which are roughly as long as the ones it was trained on. We offer a qualitative explanation of this failure and propose a novel and generic method of adding location-awareness to the attention mechanism to alleviate this issue. The new method yields a model that is robust to long inputs and achieves 18\% PER in single utterances and 20\% in 10-times longer (repeated) utterances. Finally, we propose a change to the attention mechanism that prevents it from concentrating too much on single frames, which further reduces PER to 17.6\% level.

****

Weakly-supervised Disentangling with Recurrent Transformations for 3D View Synthesis

Jimei Yang, Scott E. Reed, Ming-Hsuan Yang, Honglak Lee

An important problem for both graphics and vision is to synthesize novel views of a 3D object from a single image. This is in particular challenging due to the partial observability inherent in projecting a 3D object onto the image space, and the ill-posedness of inferring object shape and pose. However, we can train a neural network to address the problem if we restrict our attention to specific object classes (in our case faces and chairs) for which we can gather ample training data. In this paper, we propose a novel recurrent convolutional encoder-decoder network that is trained end-to-end on the task of rendering rotated objects starting from a single image. The recurrent structure allows our model to capture long- term dependencies along a sequence of transformations, and we demonstrate the quality of its predictions for human faces on the Multi-PIE dataset and for a dataset of 3D chair models, and also show its ability of disentangling latent data factors without using object class labels.

****

Backpropagation for Energy-Efficient Neuromorphic Computing

Steve K. Esser, Rathinakumar Appuswamy, Paul Merolla, John V. Arthur, Dharmendra S. Modha

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Efficient and Robust Automated Machine Learning

Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, Frank Hutter

The success of machine learning in a broad range of applications has led to an ever-growing demand for machine learning systems that can be used off the shelf by non-experts. To be effective in practice, such systems need to automatically choose a good algorithm and feature preprocessing steps for a new dataset at hand

, and also set their respective hyperparameters. Recent work has started to tackle this automated machine learning (AutoML) problem with the help of efficient Bayesian optimization methods. In this work we introduce a robust new AutoML system based on scikit-learn (using 15 classifiers, 14 feature preprocessing methods, and 4 data preprocessing methods, giving rise to a structured hypothesis space with 110 hyperparameters). This system, which we dub auto-sklearn, improves on existing AutoML methods by automatically taking into account past performance on similar datasets, and by constructing ensembles from the models evaluated during the optimization. Our system won the first phase of the ongoing ChaLearn AutoML challenge, and our comprehensive analysis on over 100 diverse datasets shows that it substantially outperforms the previous state of the art in AutoML. We also demonstrate the performance gains due to each of our contributions and derive insights into the effectiveness of the individual components of auto-sklearn.

****

Time-Sensitive Recommendation From Recurrent User Activities

Nan Du, Yichen Wang, Niao He, Jimeng Sun, Le Song

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Local Expectation Gradients for Black Box Variational Inference

Michalis Titsias RC AUEB, Miguel Lázaro-Gredilla

We introduce local expectation gradients which is a general purpose stochastic variational inference algorithm for constructing stochastic gradients by sampling from the variational distribution. This algorithm divides the problem of estimating the stochastic gradients over multiple variational parameters into smaller sub-tasks so that each sub-task explores intelligently the most relevant part of the variational distribution. This is achieved by performing an exact expectation over the single random variable that most correlates with the variational parameter of interest resulting in a Rao-Blackwellized estimate that has low variance. Our method works efficiently for both continuous and discrete random variables. Furthermore, the proposed algorithm has interesting similarities with Gibbs sampling but at the same time, unlike Gibbs sampling, can be trivially parallelized.

****

Training Restricted Boltzmann Machine via the ■Thouless-Anderson-Palmer free energy

Marylou Gabrie, Eric W. Tramel, Florent Krzakala

Restricted Boltzmann machines are undirected neural networks which have been shown tobe effective in many applications, including serving as initializations for training deep multi-layer neural networks. One of the main reasons for their success is theexistence of efficient and practical stochastic algorithms, such as contrastive divergence,for unsupervised training. We propose an alternative deterministic iterative procedure based on an improved mean field method from statistical physics known as the Thouless-Anderson-Palmer approach. We demonstrate that our algorithm provides performance equal to, and sometimes superior to, persistent contrastive divergence, while also providing a clear and easy to evaluate objective function. We believe that this strategycan be easily generalized to other models as well as to more accurate higher-order approximations, paving the way for systematic improvements in training Boltzmann machineswith hidden units.

****

High Dimensional EM Algorithm: Statistical Optimization and Asymptotic Normality

Zhaoran Wang, Quanquan Gu, Yang Ning, Han Liu

We provide a general theory of the expectation-maximization (EM) algorithm for inferring high dimensional latent variable models.   In particular, we make two contributions: (i) For parameter estimation, we propose a novel high dimensional EM algorithm which naturally incorporates sparsity structure into parameter estimation. With an appropriate initialization, this algorithm converges at a geometric rate and attains an estimator with the (near-)optimal statistical rate of c onvergence. (ii) Based on the obtained estimator, we propose a new inferential procedure for testing hypotheses for low dimensional components of high dimensional parameters. For a broad family of statistical models, our framework establishes the first computationally feasible approach for optimal estimation and asymptotic inference in high dimensions.

****

Learning Continuous Control Policies by Stochastic Value Gradients

Nicolas Heess, Gregory Wayne, David Silver, Timothy Lillicrap, Tom Erez, Yuval Tassa

We present a unified framework for learning continuous control policies usingbackpropagation. It supports stochastic control by treating stochasticity in theBellman equation as a deterministic function of exogenous noise. The productis a spectrum of general policy gradient algorithms that range from model-freemethods with value functions to model-based methods without value functions.We use learned models but only require observations from the environment insteadof observations from model-predicted trajectories, minimizing the impactof compounded model errors. We apply these algorithms first to a toy stochasticcontrol problem and then to several physics-based control problems in simulation.One of these variants, SVG(1), shows the effectiveness of learning models, valuefunctions, and policies simultaneously in continuous domains.

****

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun

State-of-the-art object detection networks depend on region proposal algorithms to hypothesize object locations. Advances like SPPnet and Fast R-CNN have reduced the running time of these detection networks, exposing region proposal computation as a bottleneck. In this work, we introduce a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. An RPN is a fully-convolutional network that simultaneously predicts object bounds and objectness scores at each position. RPNs are trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for detection. With a simple alternating optimization, RPN and Fast R-CNN can be trained to share convolutional features. For the very deep VGG-16 model, our detection system has a frame rate of 5fps (including all steps) on a GPU, while achieving state-of-the-art object detection accuracy on PASCAL VOC 2007 (73.2% mAP) and 2012 (70.4% mAP) using 300 proposals per image. Code is available at https://github.com/ShaoqingRen/faster_rcnn.

****

Efficient Non-greedy Optimization of Decision Trees

Mohammad Norouzi, Maxwell Collins, Matthew A. Johnson, David J. Fleet, Pushmeet Kohli

Decision trees and randomized forests are widely used in computer vision and machine learning. Standard algorithms for decision tree induction optimize the split functions one node at a time according to some splitting criteria. This greedy procedure often leads to suboptimal trees. In this paper, we present an algorithm for optimizing the split functions at all levels of the tree jointly with the leaf parameters, based on a global objective. We show that the problem of finding optimal linear-combination (oblique) splits for decision trees is related to structured prediction with latent variables, and we formulate a convex-concave upper bound on the tree's empirical loss. Computing the gradient of the proposed surrogate objective with respect to each training exemplar is O(d^2), where d is the tree depth, and thus training deep trees is feasible. The use of stochastic gradient descent for optimization enables effective training with large datasets. Experiments on several classification benchmarks demonstrate that the resulting non-greedy decision trees outperform greedy decision tree baselines.

****

Learning with Incremental Iterative Regularization

Lorenzo Rosasco, Silvia Villa

Within a statistical learning setting, we propose and study an iterative regularization algorithm for least squares defined by an incremental gradient method.

In particular, we show that, if all other parameters are fixed a priori, the number of passes over the data (epochs) acts as a regularization parameter, and prove strong universal consistency, i.e. almost sure convergence of the risk, as well as sharp finite sample bounds for the iterates. Our results are a step towards understanding the effect of multiple epochs in stochastic gradient techniques in machine learning and rely on integrating statistical and optimizationresults.

****

Maximum Likelihood Learning With Arbitrary Treewidth via Fast-Mixing Parameter Sets

Justin Domke

Inference is typically intractable in high-treewidth undirected graphical models, making maximum likelihood learning a challenge. One way to overcome this is to restrict parameters to a tractable set, most typically the set of tree-structured parameters. This paper explores an alternative notion of a tractable set, namely a set of “fast-mixing parameters” where Markov chain Monte Carlo (MCMC) inference can be guaranteed to quickly converge to the stationary distribution. While it is common in practice to approximate the likelihood gradient using samples obtained from MCMC, such procedures lack theoretical guarantees. This paper proves that for any exponential family with bounded sufficient statistics, (not just graphical models) when parameters are constrained to a fast-mixing set, gradient descent with gradients approximated by sampling will approximate the maximum likelihood solution inside the set with high-probability. When unregularized, to find a solution epsilon-accurate in log-likelihood requires a total amount of effort cubic in 1/epsilon, disregarding logarithmic factors. When ridge-regularized, strong convexity allows a solution epsilon-accurate in parameter distance with an effort quadratic in 1/epsilon. Both of these provide of a fully-polynomial time randomized approximation scheme.

****

Sampling from Probabilistic Submodular Models

Alkis Gotovos, Hamed Hassani, Andreas Krause

Submodular and supermodular functions have found wide applicability in machine learning, capturing notions such as diversity and regularity, respectively. These notions have deep consequences for optimization, and the problem of (approximately) optimizing submodular functions has received much attention. However, beyond optimization, these notions allow specifying expressive probabilistic models that can be used to quantify predictive uncertainty via marginal inference. Prominent, well-studied special cases include Ising models and determinantal point processes, but the general class of log-submodular and log-supermodular models is much richer and little studied. In this paper, we investigate the use of Markov chain Monte Carlo sampling to perform approximate inference in general log-submodular and log-supermodular models. In particular, we consider a simple Gibbs sampling procedure, and establish two sufficient conditions, the first guaranteeing polynomial-time, and the second fast (O(nlogn)) mixing. We also evaluate the efficiency of the Gibbs sampler on three examples of such models, and compare against a recently proposed variational approach.

****

A class of network models recoverable by spectral clustering

Yali Wan, Marina Meila

Finding communities in networks is a problem that remains difficult, in spite of the amount of attention it has recently received. The Stochastic Block-Model (SBM) is a generative model for graphs with communities for which, because of its simplicity, the theoretical understanding has advanced fast in recent years. In particular, there have been various results showing that simple versions of spectralclustering using the Normalized Laplacian of the graph can recoverthe communities almost perfectly with high probability. Here we show that essentially the same algorithm used for the SBM and for its extension called Degree-Corrected SBM, works on a wider class of Block-Models, which we call Preference Frame Models, with essentially the same guarantees. Moreover, the parametrization we introduce clearly exhibits the free parameters needed to specify this class of models, and results in bounds that expose with more clarity the parameters that control the recovery error in this model class.

****

Closed-form Estimators for High-dimensional Generalized Linear Models

Eunho Yang, Aurelie C. Lozano, Pradeep K. Ravikumar

We propose a class of closed-form estimators for GLMs under high-dimensional sampling regimes. Our class of estimators is based on deriving closed-form variants of the vanilla unregularized MLE but which are (a) well-defined even under high-dimensional settings, and (b) available in closed-form. We then perform thresholding operations on this MLE variant to obtain our class of estimators. We derive a unified statistical analysis of our class of estimators, and show that it enjoys strong statistical guarantees in both parameter error as well as variable selection, that surprisingly match those of the more complex regularized GLM MLEs, even while our closed-form estimators are computationally much simpler. We derive instantiations of our class of closed-form estimators, as well as corollaries of our general theorem, for the special cases of logistic, exponential and Poisson regression models. We corroborate the surprising statistical and computational performance of our class of estimators via extensive simulations.

****

Expressing an Image Stream with a Sequence of Natural Sentences

Cesc C. Park, Gunhee Kim

We propose an approach for generating a sequence of natural sentences for an image stream. Since general users usually take a series of pictures on their special moments, much online visual information exists in the form of image streams, for which it would better take into consideration of the whole set to generate natural language descriptions. While almost all previous studies have dealt with the relation between a single image and a single natural sentence, our work extends both input and output dimension to a sequence of images and a sequence of sentences. To this end, we design a novel architecture called coherent recurrent convolutional network (CRCN), which consists of convolutional networks, bidirectional recurrent networks, and entity-based local coherence model. Our approach directly learns from vast user-generated resource of blog posts as text-image parallel training data. We demonstrate that our approach outperforms other state-of-the-art candidate methods, using both quantitative measures (e.g. BLEU and top-K recall) and user studies via Amazon Mechanical Turk.

****

Learning spatiotemporal trajectories from manifold-valued longitudinal data

Jean-Baptiste SCHIRATTI, Stéphanie ALLASSONNIERE, Olivier Colliot, Stanley DURRLEMAN

We propose a Bayesian mixed-effects model to learn typical scenarios of changes from longitudinal manifold-valued data, namely repeated measurements of the same objects or individuals at several points in time. The model allows to estimate a group-average trajectory in the space of measurements. Random variations of this trajectory result from spatiotemporal transformations, which allow changes in the direction of the trajectory and in the pace at which trajectories are followed. The use of the tools of Riemannian geometry allows to derive a generic algorithm for any kind of data with smooth constraints, which lie therefore on a Riemannian manifold. Stochastic approximations of the Expectation-Maximization algorithm is used to estimate the model parameters in this highly non-linear setting.The method is used to estimate a data-driven model of the progressive impairments of cognitive functions during the onset of Alzheimer's disease. Experimental results show that the model correctly put into correspondence the age at which each individual was diagnosed with the disease, thus validating the fact that it effectively estimated a normative scenario of disease progression. Random effects provide unique insights into the variations in the ordering and timing of the succession of cognitive impairments across different individuals.

****

Fast Classification Rates for High-dimensional Gaussian Generative Models

Tianyang Li, Adarsh Prasad, Pradeep K. Ravikumar

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Adaptive Online Learning

Dylan J. Foster, Alexander Rakhlin, Karthik Sridharan

We propose a general framework for studying adaptive regret bounds in the online learning setting, subsuming model selection and data-dependent bounds. Given a data- or model-dependent bound we ask, “Does there exist some algorithm achieving this bound?” We show that modifications to recently introduced sequential complexity measures can be used to answer this question by providing sufficient conditions under which adaptive rates can be achieved. In particular each adaptive rate induces a set of so-called offset complexity measures, and obtaining small upper bounds on these quantities is sufficient to demonstrate achievability. A cornerstone of our analysis technique is the use of one-sided tail inequalities to bound suprema of offset random processes.Our framework recovers and improves a wide variety of adaptive bounds including quantile bounds, second order data-dependent bounds, and small loss bounds. In addition we derive a new type of adaptive bound for online linear optimization based on the spectral norm, as well as a new online PAC-Bayes theorem.

****

Robust Regression via Hard Thresholding

Kush Bhatia, Prateek Jain, Purushottam Kar

We study the problem of Robust Least Squares Regression (RLSR) where several response variables can be adversarially corrupted. More specifically, for a data matrix X \in \R^{p x n} and an underlying model w, the response vector is generated as y = X'w + b where b \in n is the corruption vector supported over at most C.n coordinates. Existing exact recovery results for RLSR focus solely on L1-penalty based convex formulations and impose relatively strict model assumptions such as requiring the corruptions b to be selected independently of X.In this work, we study a simple hard-thresholding algorithm called TORRENT which, under mild conditions on X, can recover w* exactly even if b corrupts the response variables in an adversarial manner, i.e. both the support and entries of b are selected adversarially after observing X and w. Our results hold under deterministic assumptions which are satisfied if X is sampled from any sub-Gaussian distribution. Finally unlike existing results that apply only to a fixed w, generated independently of X, our results are universal and hold for any w* \in \R^p.Next, we propose gradient descent-based extensions of TORRENT that can scale efficiently to large scale problems, such as high dimensional sparse recovery. and prove similar recovery guarantees for these extensions. Empirically we find TORRENT, and more so its extensions, offering significantly faster recovery than the state-of-the-art L1 solvers. For instance, even on moderate-sized datasets (with p = 50K) with around 40% corrupted responses, a variant of our proposed method called TORRENT-HYB is more than 20x faster than the best L1 solver.

****

b-bit Marginal Regression

Martin Slawski, Ping Li

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Spectral Norm Regularization of Orthonormal Representations for Graph Transduction

Rakesh Shivanna, Bibaswan K. Chatterjee, Raman Sankaran, Chiranjib Bhattacharyya, Francis Bach

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Randomized Block Krylov Methods for Stronger and Faster Approximate Singular Value Decomposition

Cameron Musco, Christopher Musco

Since being analyzed by Rokhlin, Szlam, and Tygert and popularized by Halko, Martinsson, and Tropp, randomized Simultaneous Power Iteration has become the method of choice for approximate singular value decomposition. It is more accurate than simpler sketching algorithms, yet still converges quickly for any matrix, independently of singular value gaps. After ~O(1/epsilon) iterations, it gives a low-rank approximation within (1+epsilon) of optimal for spectral norm error.We give the first provable runtime improvement on Simultaneous Iteration: a randomized block Krylov method, closely related to the classic Block Lanczos algorithm, gives the same guarantees in just ~O(1/sqrt(epsilon)) iterations and performs substantially better experimentally. Our analysis is the first of a Krylov subspace method that does not depend on singular value gaps, which are unreliable in practice.Furthermore, while it is a simple accuracy benchmark, even (1+epsilon) error for spectral norm low-rank approximation does not imply that an algorithm returns high quality principal components, a major issue for data applications. We address this problem for the first time by showing that both Block Krylov Iteration and Simultaneous Iteration give nearly optimal PCA for any matrix. This result further justifies their strength over non-iterative sketching methods.

****

Optimal Testing for Properties of Distributions

Jayadev Acharya, Constantinos Daskalakis, Gautam Kamath

Given samples from an unknown distribution, p, is it possible to distinguish whether p belongs to some class of distributions C versus p being far from every distribution in C? This fundamental question has receivedtremendous attention in Statistics, albeit focusing onasymptotic analysis, as well as in Computer Science, wherethe emphasis has been on small sample size and computationalcomplexity. Nevertheless, even for basic classes ofdistributions such as monotone, log-concave, unimodal, and monotone hazard rate, the optimal sample complexity is unknown.We provide a general approach via which we obtain sample-optimal and computationally efficient testers for all these distribution families. At the core of our approach is an algorithm which solves the following problem:Given samplesfrom an unknown distribution p, and a known distribution q, are p and q close in Chi^2-distance, or far in total variation distance?The optimality of all testers is established by providing matching lower bounds. Finally, a necessary building block for our tester and important byproduct of our work are the first known computationally efficient proper learners for discretelog-concave and monotone hazard rate distributions. We exhibit the efficacy of our testers via experimental analysis.

****

Combinatorial Cascading Bandits

Branislav Kveton, Zheng Wen, Azin Ashkan, Csaba Szepesvari

We propose combinatorial cascading bandits, a class of partial monitoring problems where at each step a learning agent chooses a tuple of ground items subject to constraints and receives a reward if and only if the weights of all chosen items are one. The weights of the items are binary, stochastic, and drawn independently of each other. The agent observes the index of the first chosen item whose weight is zero. This observation model arises in network routing, for instance, where the learning agent may only observe the first link in the routing path which is down, and blocks the path. We propose a UCB-like algorithm for solving our problems, CombCascade; and prove gap-dependent and gap-free upper bounds on its n-step regret. Our proofs build on recent work in stochastic combinatorial semi-bandits but also address two novel challenges of our setting, a non-linear reward function and partial observability. We evaluate CombCascade on two real-world problems and show that it performs well even when our modeling assumptions are violated. We also demonstrate that our setting requires a new learning algorithm.

****

Probabilistic Curve Learning: Coulomb Repulsion and the Electrostatic Gaussian Process

Ye Wang, David B. Dunson

Learning of low dimensional structure in multidimensional data is a canonical problem in machine learning. One common approach is to suppose that the observed data are close to a lower-dimensional smooth manifold. There are a rich variety of manifold learning methods available, which allow mapping of data points to the manifold. However, there is a clear lack of probabilistic methods that allow learning of the manifold along with the generative distribution of the observed data. The best attempt is the Gaussian process latent variable model (GP-LVM), but identifiability issues lead to poor performance. We solve these issues by proposing a novel Coulomb repulsive process (Corp) for locations of points on the manifold, inspired by physical models of electrostatic interactions among particles. Combining this process with a GP prior for the mapping function yields a novel electrostatic GP (electroGP) process. Focusing on the simple case of a one-dimensional manifold, we develop efficient inference algorithms, and illustrate substantially improved performance in a variety of experiments including filling in missing frames in video.

****

Training Very Deep Networks

Rupesh K. Srivastava, Klaus Greff, Jürgen Schmidhuber

Theoretical and empirical evidence indicates that the depth of neural networks is crucial for their success. However, training becomes more difficult as depth increases, and training of very deep networks remains an open problem. Here we introduce a new architecture designed to overcome this. Our so-called highway networks allow unimpeded information flow across many layers on information highways. They are inspired by Long Short-Term Memory recurrent networks and use adaptive gating units to regulate the information flow. Even with hundreds of layers, highway networks can be trained directly through simple gradient descent. This enables the study of extremely deep and efficient architectures.

****

Fast and Memory Optimal Low-Rank Matrix Approximation

Se-Young Yun, marc lelarge, Alexandre Proutiere

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Character-level Convolutional Networks for Text Classification

Xiang Zhang, Junbo Zhao, Yann LeCun

This article offers an empirical exploration on the use of character-level convolutional networks (ConvNets) for text classification. We constructed several large-scale datasets to show that character-level convolutional networks could achieve state-of-the-art or competitive results. Comparisons are offered against traditional models such as bag of words, n-grams and their TFIDF variants, and deep learning models such as word-based ConvNets and recurrent neural networks.

****

Interactive Control of Diverse Complex Characters with Neural Networks

Igor Mordatch, Kendall Lowrey, Galen Andrew, Zoran Popovic, Emanuel V. Todorov

We present a method for training recurrent neural networks to act as near-optimal feedback controllers. It is able to generate stable and realistic behaviors for a range of dynamical systems and tasks -- swimming, flying, biped and quadruped walking with different body morphologies. It does not require motion capture or task-specific features or state machines. The controller is a neural network, having a large number of feed-forward units that learn elaborate state-action mappings, and a small number of recurrent units that implement memory states beyond the physical system state. The action generated by the network is defined as velocity. Thus the network is not learning a control policy, but rather the dynamics under an implicit policy. Essential features of the method include interleaving supervised learning with trajectory optimization, injecting noise during tra ining, training for unexpected changes in the task specification, and using the trajectory optimizer to obtain optimal feedback gains in addition to optimal actions.

****

Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets

Armand Joulin, Tomas Mikolov

Despite the recent achievements in machine learning, we are still very far from achieving real artificial intelligence. In this paper, we discuss the limitations of standard deep learning approaches and show that some of these limitations can be overcome by learning how to grow the complexity of a model in a structured way. Specifically, we study the simplest sequence prediction problems that are beyond the scope of what is learnable with standard recurrent networks, algorithmically generated sequences which can only be learned by models which have the capacity to count and to memorize sequences. We show that some basic algorithms can be learned from sequential data using a recurrent network associated with a trainable memory.

****

Grammar as a Foreign Language

Oriol Vinyals, ■ukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, Geoffrey Hinton

Syntactic constituency parsing is a fundamental problem in naturallanguage processing which has been the subject of intensive researchand engineering for decades. As a result, the most accurate parsersare domain specific, complex, and inefficient. In this paper we showthat the domain agnostic attention-enhanced sequence-to-sequence modelachieves state-of-the-art results on the most widely used syntacticconstituency parsing dataset, when trained on a large synthetic corpusthat was annotated using existing parsers. It also matches theperformance of standard parsers when trained on a smallhuman-annotated dataset, which shows that this model is highlydata-efficient, in contrast to sequence-to-sequence models without theattention mechanism. Our parser is also fast, processing over ahundred sentences per second with an unoptimized CPU implementation.

****

Practical and Optimal LSH for Angular Distance

Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya Razenshteyn, Ludwig Schmidt

We show the existence of a Locality-Sensitive Hashing (LSH) family for the angular distance that yields an approximate Near Neighbor Search algorithm with the asymptotically optimal running time exponent. Unlike earlier algorithms with this property (e.g., Spherical LSH (Andoni-Indyk-Nguyen-Razenshteyn 2014) (Andoni-Razenshteyn 2015)), our algorithm is also practical, improving upon the well-studied hyperplane LSH (Charikar 2002) in practice. We also introduce a multiprobe version of this algorithm and conduct an experimental evaluation on real and synthetic data sets.We complement the above positive results with a fine-grained lower bound for the quality of any LSH family for angular distance. Our lower bound implies that the above LSH family exhibits a trade-off between evaluation time and quality that is close to optimal for a natural class of LSH functions.

****

GP Kernels for Cross-Spectrum Analysis

Kyle R. Ulrich, David E. Carlson, Kafui Dzirasa, Lawrence Carin

Multi-output Gaussian processes provide a convenient framework for multi-task problems. An illustrative and motivating example of a multi-task problem is multi-region electrophysiological time-series data, where experimentalists are interested in both power and phase coherence between channels. Recently, Wilson and Adams (2013) proposed the spectral mixture (SM) kernel to model the spectral density of a single task in a Gaussian process framework. In this paper, we develop a novel covariance kernel for multiple outputs, called the cross-spectral mixture (CSM) kernel. This new, flexible kernel represents both the power and phase relationship between multiple observation channels. We demonstrate the expressive capabilities of the CSM kernel through implementation of a Bayesian hidden Markov model, where the emission distribution is a multi-output Gaussian process with a CSM covariance kernel. Results are presented for measured multi-region el ectrophysiological data.

****

A Framework for Individualizing Predictions of Disease Trajectories by Exploiting Multi-Resolution Structure

Peter Schulam, Suchi Saria

For many complex diseases, there is a wide variety of ways in which an individual can manifest the disease. The challenge of personalized medicine is to develop tools that can accurately predict the trajectory of an individual's disease, which can in turn enable clinicians to optimize treatments. We represent an individual's disease trajectory as a continuous-valued continuous-time function describing the severity of the disease over time. We propose a hierarchical latent variable model that individualizes predictions of disease trajectories. This model shares statistical strength across observations at different resolutions--the population, subpopulation and the individual level. We describe an algorithm for learning population and subpopulation parameters offline, and an online procedure for dynamically learning individual-specific parameters. Finally, we validate our model on the task of predicting the course of interstitial lung disease, a leading cause of death among patients with the autoimmune disease scleroderma. We compare our approach against state-of-the-art and demonstrate significant improvements in predictive accuracy.

****

Local Smoothness in Variance Reduced Optimization

Daniel Vainsencher, Han Liu, Tong Zhang

Abstract We propose a family of non-uniform sampling strategies to provably speed up a class of stochastic optimization algorithms with linear convergence including Stochastic Variance Reduced Gradient (SVRG) and Stochastic Dual Coordinate Ascent (SDCA). For a large family of penalized empirical risk minimization problems, our methods exploit data dependent local smoothness of the loss functions near the optimum, while maintaining convergence guarantees. Our bounds are the first to quantify the advantage gained from local smoothness which are significant for some problems significantly better. Empirically, we provide thorough numerical results to back up our theory. Additionally we present algorithms exploiting local smoothness in more aggressive ways, which perform even better in practice.

****

Unlocking neural population non-stationarities using hierarchical dynamics models

Mijung Park, Gergo Bohner, Jakob H. Macke

Neural population activity often exhibits rich variability. This variability is thought to arise from single-neuron stochasticity, neural dynamics on short time-scales, as well as from modulations of neural firing properties on long time-scales, often referred to as non-stationarity. To better understand the nature of co-variability in neural circuits and their impact on cortical information processing, we introduce a hierarchical dynamics model that is able to capture inter-trial modulations in firing rates, as well as neural population dynamics. We derive an algorithm for Bayesian Laplace propagation for fast posterior inference, and demonstrate that our model provides a better account of the structure of neural firing than existing stationary dynamics models, when applied to neural population recordings from primary visual cortex.

****

Pointer Networks

Oriol Vinyals, Meire Fortunato, Navdeep Jaitly

We introduce a new neural architecture to learn the conditional probability of an output sequence with elements that arediscrete tokens corresponding to positions in an input sequence.Such problems cannot be trivially addressed by existent approaches such as sequence-to-sequence and Neural Turing Machines,because the number of target classes in eachstep of the output depends on the length of the input, which is variable.Problems such as sorting variable sized sequences, and various combinatorialoptimization problems belong to this class. Our model solvesthe problem of variable size output dictionaries using a recently proposedmecha nism of neural attention. It differs from the previous attentionattempts in that, instead of using attention to blend hidden units of anencoder to a context vector at each decoder step, it uses attention asa pointer to select a member of the input sequence as the output. We call this architecture a Pointer Net (Ptr-Net).We show Ptr-Nets can be used to learn approximate solutions to threechallenging geometric problems -- finding planar convex hulls, computingDelaunay triangulations, and the planar Travelling Salesman Problem-- using training examples alone. Ptr-Nets not only improve oversequence-to-sequence with input attention, butalso allow us to generalize to variable size output dictionaries.We show that the learnt models generalize beyond the maximum lengthsthey were trained on. We hope our results on these taskswill encourage a broader exploration of neural learning for discreteproblems.

****

Fast and Accurate Inference of Plackett–Luce Models

Lucas Maystre, Matthias Grossglauser

We show that the maximum-likelihood (ML) estimate of models derived from Luce's choice axiom (e.g., the Plackett-Luce model) can be expressed as the stationary distribution of a Markov chain. This conveys insight into several recently proposed spectral inference algorithms. We take advantage of this perspective and formulate a new spectral algorithm that is significantly more accurate than previous ones for the Plackett--Luce model. With a simple adaptation, this algorithm can be used iteratively, producing a sequence of estimates that converges to the ML estimate. The ML version runs faster than competing approaches on a benchmark of five datasets. Our algorithms are easy to implement, making them relevant for practitioners at large.

****

Learning Bayesian Networks with Thousands of Variables

Mauro Scanagatta, Cassio P. de Campos, Giorgio Corani, Marco Zaffalon

We present a method for learning Bayesian networks from data sets containingthousands of variables without the need for structure constraints. Our approachis made of two parts. The first is a novel algorithm that effectively explores thespace of possible parent sets of a node. It guides the exploration towards themost promising parent sets on the basis of an approximated score function thatis computed in constant time. The second part is an improvement of an existingordering-based algorithm for structure optimization. The new algorithm provablyachieves a higher score compared to its original formulation. On very large datasets containing up to ten thousand nodes, our novel approach consistently outper-forms the state of the art.

****

Differentially Private Learning of Structured Discrete Distributions

Ilias Diakonikolas, Moritz Hardt, Ludwig Schmidt

We investigate the problem of learning an unknown probability distribution over a discrete population from random samples. Our goal is to design efficient algorithms that simultaneously achieve low error in total variation norm while guaranteeing Differential Privacy to the individuals of the population.We describe a general approach that yields near sample-optimal and computationally efficient differentially private estimators for a wide range of well-studied and natural distribution families. Our theoretical results show that for a wide variety of structured distributions there exist private estimation algorithms that are nearly as efficient - both in terms of sample size and running time - as their non-private counterparts. We complement our theoretical guarantees with an experimental evaluation. Our experiments illustrate the speed and accuracy of our private estimators on both synthetic mixture models and a large public data set.

****

Generative Image Modeling Using Spatial LSTMs

Lucas Theis, Matthias Bethge

Modeling the distribution of natural images is challenging, partly because of strong statistical dependencies which can extend over hundreds of pixels. Recurrent neural networks have been successful in capturing long-range dependencies in a number of problems but only recently have found their way into generative image

models. We here introduce a recurrent image model based on multi-dimensional long short-term memory units which are particularly suited for image modeling due to their spatial structure. Our model scales to images of arbitrary size and its likelihood is computationally tractable. We find that it outperforms the state of the art in quantitative comparisons on several image datasets and produces promising results when used for texture synthesis and inpainting.
****
Sparse PCA via Bipartite Matchings
Megasthenis Asteris, Dimitris Papailiopoulos, Anastasios Kyrillidis, Alexandros G. Dimakis
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Market Scoring Rules Act As Opinion Pools For Risk-Averse Agents
Mithun Chakraborty, Sanmay Das
A market scoring rule (MSR) – a popular tool for designing algorithmic prediction markets – is an incentive-compatible mechanism for the aggregation of probabilistic beliefs from myopic risk-neutral agents. In this paper, we add to a growing body of research aimed at understanding the precise manner in which the price process induced by a MSR incorporates private information from agents who deviate from the assumption of risk-neutrality. We first establish that, for a myopic trading agent with a risk-averse utility function, a MSR satisfying mild regularity conditions elicits the agent’s risk-neutral probability conditional on the latest market state rather than her true subjective probability. Hence, we show that a MSR under these conditions effectively behaves like a more traditional method of belief aggregation, namely an opinion pool, for agents’ true probabilities. In particular, the logarithmic market scoring rule acts as a logarithmic pool for constant absolute risk aversion utility agents, and as a linear pool for an atypical budget-constrained agent utility with decreasing absolute risk aversion. We also point out the interpretation of a market maker under these conditions as a Bayesian learner even when agent beliefs are static.
****
Lifted Inference Rules With Constraints
Happy Mittal, Anuj Mahajan, Vibhav G. Gogate, Parag Singla
Lifted inference rules exploit symmetries for fast reasoning in statistical rela-tional models. Computational complexity of these rules is highly dependent onthe choice of the constraint language they operate on and therefore coming upwith the right kind of representation is critical to the success of lifted inference. In this paper, we propose a new constraint language, called setineq, which allowssubset, equality and inequality constraints, to represent substitutions over the vari-ables in the theory. Our constraint formulation is strictly more expressive thanexisting representations, yet easy to operate on. We reformulate the three mainlifting rules: decomposer, generalized binomial and the recently proposed singleoccurrence for MAP inference, to work with our constraint representation. Exper-iments on benchmark MLNs for exact and sampling based inference demonstratethe effectiveness of our approach over several other existing techniques.
****
LASSO with Non-linear Measurements is Equivalent to One With Linear Measurements
CHRISTOS THRAMPOULIDIS, Ehsan Abbasi, Babak Hassibi
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Natural Neural Networks
Guillaume Desjardins, Karen Simonyan, Razvan Pascanu, koray kavukcuoglu
We introduce Natural Neural Networks, a novel family of algorithms that speed up convergence by adapting their internal representation during training to improv

e conditioning of the Fisher matrix. In particular, we show a specific example that employs a simple and efficient reparametrization of the neural network weights by implicitly whitening the representation obtained at each layer, while preserving the feed-forward computation of the network. Such networks can be trained efficiently via the proposed Projected Natural Gradient Descent algorithm (PRONG), which amortizes the cost of these reparametrizations over many parameter updates and is closely related to the Mirror Descent online learning algorithm. We highlight the benefits of our method on both unsupervised and supervised learning tasks, and showcase its scalability by training on the large-scale ImageNet Challenge dataset.

****

Scalable Adaptation of State Complexity for Nonparametric Hidden Markov Models
Michael C. Hughes, William T. Stephenson, Erik Sudderth

Bayesian nonparametric hidden Markov models are typically learned via fixed truncations of the infinite state space or local Monte Carlo proposals that make small changes to the state space. We develop an inference algorithm for the sticky hierarchical Dirichlet process hidden Markov model that scales to big datasets by processing a few sequences at a time yet allows rapid adaptation of the state space cardinality. Unlike previous point-estimate methods, our novel variational bound penalizes redundant or irrelevant states and thus enables optimization of the state space. Our birth proposals use observed data statistics to create useful new states that escape local optima. Merge and delete proposals remove ineffective states to yield simpler models with more affordable future computations. Experiments on speaker diarization, motion capture, and epigenetic chromatin datasets discover models that are more compact, more interpretable, and better aligned to ground truth segmentations than competitors. We have released an open-source Python implementation which can parallelize local inference steps across sequences.

****

Inference for determinantal point processes without spectral knowledge
Rémi Bardenet, Michalis Titsias RC AUEB

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

A Bayesian Framework for Modeling Confidence in Perceptual Decision Making
Koosha Khalvati, Rajesh PN Rao

The degree of confidence in one's choice or decision is a critical aspect of perceptual decision making. Attempts to quantify a decision maker's confidence by measuring accuracy in a task have yielded limited success because confidence and accuracy are typically not equal. In this paper, we introduce a Bayesian framework to model confidence in perceptual decision making. We show that this model, based on partially observable Markov decision processes (POMDPs), is able to predict confidence of a decision maker based only on the data available to the experimenter. We test our model on two experiments on confidence-based decision making involving the well-known random dots motion discrimination task. In both experiments, we show that our model's predictions closely match experimental data. Additionally, our model is also consistent with other phenomena such as the hard-easy effect in perceptual decision making.

****

Sample Complexity of Episodic Fixed-Horizon Reinforcement Learning
Christoph Dann, Emma Brunskill

Recently, there has been significant progress in understanding reinforcement learning in discounted infinite-horizon Markov decision processes (MDPs) by deriving tight sample complexity bounds. However, in many real-world applications, an interactive learning agent operates for a fixed or bounded period of time, for example tutoring students for exams or handling customer service requests. Such scenarios can often be better treated as episodic fixed-horizon MDPs, for which only looser bounds on the sample complexity exist. A natural notion of sample com plexity in this setting is the number of episodes required to guarantee a certain performance with high probability (PAC guarantee). In this paper, we derive an upper PAC bound of order $O(|S|^2|A|H^2 \log(1/\delta)/\blacksquare^2)$ and a lower PAC bound $\Omega(|S||A|H^2 \log(1/(\delta+c))/\blacksquare^2)$ (ignoring log-terms) that match up to log-terms and an additional linear dependency on the number of states |S|. The lower bound is the first of its kind for this setting. Our upper bound leverages Bernstein's inequality to improve on previous bounds for episodic finite-horizon MDPs which have a time-horizon dependency of at least H^3.

****

Algorithms with Logarithmic or Sublinear Regret for Constrained Contextual Bandits
Huasen Wu, R. Srikant, Xin Liu, Chong Jiang
We study contextual bandits with budget and time constraints under discrete contexts, referred to as constrained contextual bandits. The time and budget constraints significantly complicate the exploration and exploitation tradeoff because they introduce complex coupling among contexts over time. To gain insight, we first study unit-cost systems with known context distribution. When the expected rewards are known, we develop an approximation of the oracle, referred to Adaptive-Linear-Programming(ALP), which achieves near-optimality and only requires the ordering of expected rewards. With these highly desirable features, we then combine ALP with the upper-confidence-bound (UCB) method in the general case where the expected rewards are unknown a priori. We show that the proposed UCB-ALP algorithm achieves logarithmic regret except in certain boundary cases.Further, we design algorithms and obtain similar regret analysis results for more general systems with unknown context distribution or heterogeneous costs. To the best of our knowledge, this is the first work that shows how to achieve logarithmic regret in constrained contextual bandits. Moreover, this work also sheds light on the study of computationally efficient algorithms for general constrained contextual bandits.

****

Latent Bayesian melding for integrating individual and population models
Mingjun Zhong, Nigel Goddard, Charles Sutton
In many statistical problems, a more coarse-grained model may be suitable for population-level behaviour, whereas a more detailed model is appropriate for accurate modelling of individual behaviour. This raises the question of how to integrate both types of models. Methods such as posterior regularization follow the idea of generalized moment matching, in that they allow matchingexpectations between two models, but sometimes both models are most conveniently expressed as latent variable models. We propose latent Bayesian melding, which is motivated by averaging the distributions over populations statistics of both the individual-level and the population-level models under a logarithmic opinion pool framework. In a case study on electricity disaggregation, which is a type of single-channel blind source separation problem, we show that latent Bayesian melding leads to significantly more accurate predictions than an approach based solely on generalized moment matching.

****

Regressive Virtual Metric Learning
Michaël Perrot, Amaury Habrard
We are interested in supervised metric learning of Mahalanobis like distances. Existing approaches mainly focus on learning a new distance using similarity and dissimilarity constraints between examples. In this paper, instead of bringing closer examples of the same class and pushing far away examples of different classes we propose to move the examples with respect to virtual points. Hence, each example is brought closer to a a priori defined virtual point reducing the number of constraints to satisfy. We show that our approach admits a closed form solution which can be kernelized. We provide a theoretical analysis showing the consistency of the approach and establishing some links with other classical metric learning methods. Furthermore we propose an efficient solution to the difficult problem of selecting virtual points based in part on recent works in optimal transport. Lastly, we evaluate our approach on several state of the art datasets.

****

Halting in Random Walk Kernels

Mahito Sugiyama, Karsten Borgwardt

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Kullback-Leibler Proximal Variational Inference

Mohammad Emtiyaz E. Khan, Pierre Baque, François Fleuret, Pascal Fua

We propose a new variational inference method based on the Kullback-Leibler (KL) proximal term. We make two contributions towards improving efficiency of variational inference. Firstly, we derive a KL proximal-point algorithm and show its equivalence to gradient descent with natural gradient in stochastic variational inference. Secondly, we use the proximal framework to derive efficient variational algorithms for non-conjugate models. We propose a splitting procedure to separate non-conjugate terms from conjugate ones. We then linearize the non-conjugate terms and show that the resulting subproblem admits a closed-form solution. Overall, our approach converts a non-conjugate model to subproblems that involve inference in well-known conjugate models. We apply our method to many models and derive generalizations for non-conjugate exponential family. Applications to real-world datasets show that our proposed algorithms are easy to implement, fast to converge, perform well, and reduce computations.

****

A Convergent Gradient Descent Algorithm for Rank Minimization and Semidefinite Programming from Random Linear Measurements

Qinqing Zheng, John Lafferty

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

On-the-Job Learning with Bayesian Decision Theory

Keenon Werling, Arun Tejasvi Chaganty, Percy S. Liang, Christopher D. Manning

Our goal is to deploy a high-accuracy system starting with zero training examples. We consider an “on-the-job” setting, where as inputs arrive, we use real-time crowdsourcing to resolve uncertainty where needed and output our prediction when confident. As the model improves over time, the reliance on crowdsourcing queries decreases. We cast our setting as a stochastic game based on Bayesian decision theory, which allows us to balance latency, cost, and accuracy objectives in a principled way. Computing the optimal policy is intractable, so we develop an approximation based on Monte Carlo Tree Search. We tested our approach on three datasets-- named-entity recognition, sentiment classification, and image classification. On the NER task we obtained more than an order of magnitude reduction in cost compared to full human annotation, while boosting performance relative to the expert provided labels. We also achieve a 8% F1 improvement over having a single human label the whole set, and a 28% F1 improvement over online learning.

****

Spatial Transformer Networks

Max Jaderberg, Karen Simonyan, Andrew Zisserman, koray kavukcuoglu

Convolutional Neural Networks define an exceptionallypowerful class of model, but are still limited by the lack of abilityto be spatially invariant to the input data in a computationally and parameterefficient manner. In this work we introduce a new learnable module, theSpatial Transformer, which explicitly allows the spatial manipulation ofdata within the network. This differentiable module can be insertedinto existing convolutional architectures, giving neural networks the ability toactively spatially transform feature maps, conditional on the feature map itself,without any extra training supervision or modification to the optimisation process. We show that the useof spatial transformers results in models which learn invariance to translation,scale, rotation and more generic warping, res ulting in state-of-the-artperformance on several benchmarks, and for a numberof classes of transformations.

****

Precision-Recall-Gain Curves: PR Analysis Done Right

Peter Flach, Meelis Kull

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Planar Ultrametrics for Image Segmentation

Julian E. Yarkony, Charless Fowlkes

We study the problem of hierarchical clustering on planar graphs. We formulate this in terms of finding the closest ultrametric to a specified set of distances and solve it using an LP relaxation that leverages minimum cost perfect matching as a subroutine to efficiently explore the space of planar partitions. We apply our algorithm to the problem of hierarchical image segmentation.

****

Sparse Local Embeddings for Extreme Multi-label Classification

Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, Prateek Jain

The objective in extreme multi-label learning is to train a classifier that can automatically tag a novel data point with the most relevant subset of labels from an extremely large label set. Embedding based approaches make training and prediction tractable by assuming that the training label matrix is low-rank and hence the effective number of labels can be reduced by projecting the high dimensional label vectors onto a low dimensional linear subspace. Still, leading embedding approaches have been unable to deliver high prediction accuracies or scale to large problems as the low rank assumption is violated in most real world applications.This paper develops the SLEEC classifier to address both limitations. The main technical contribution in SLEEC is a formulation for learning a small ensemble of local distance preserving embeddings which can accurately predict infrequently occurring (tail) labels. This allows SLEEC to break free of the traditional low-rank assumption and boost classification accuracy by learning embeddings which preserve pairwise distances between only the nearest label vectors. We conducted extensive experiments on several real-world as well as benchmark data sets and compare our method against state-of-the-art methods for extreme multi-label classification. Experiments reveal that SLEEC can make significantly more accurate predictions then the state-of-the-art methods including both embeddings (by as much as 35%) as well as trees (by as much as 6%). SLEEC can also scale efficiently to data sets with a million labels which are beyond the pale of leading embedding methods.

****

Super-Resolution Off the Grid

Qingqing Huang, Sham M. Kakade

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Automatic Variational Inference in Stan

Alp Kucukelbir, Rajesh Ranganath, Andrew Gelman, David Blei

Variational inference is a scalable technique for approximate Bayesian inference. Deriving variational inference algorithms requires tedious model-specific calculations; this makes it difficult for non-experts to use. We propose an automatic variational inference algorithm, automatic differentiation variational inference (ADVI); we implement it in Stan (code available), a probabilistic programming system. In ADVI the user provides a Bayesian model and a dataset, nothing else. We make no conjugacy assumptions and support a broad class of models. The algorithm automatically determines an appropriate variational family and optimizes the variational objective. We compare ADVI to MCMC sampling across hierarchica

l generalized linear models, nonconjugate matrix factorization, and a mixture model. We train the mixture model on a quarter million images. With ADVI we can use variational inference on any model we write in Stan.
****

Extending Gossip Algorithms to Distributed Estimation of U-statistics
Igor Colin, Aurélien Bellet, Joseph Salmon, Stéphan Clémençon
Efficient and robust algorithms for decentralized estimation in networks are essential to many distributed systems. Whereas distributed estimation of sample mean statistics has been the subject of a good deal of attention, computation of U-statistics, relying on more expensive averaging over pairs of observations, is a less investigated area. Yet, such data functionals are essential to describe global properties of a statistical population, with important examples including Area Under the Curve, empirical variance, Gini mean difference and within-cluster point scatter. This paper proposes new synchronous and asynchronous randomized gossip algorithms which simultaneously propagate data across the network and maintain local estimates of the U-statistic of interest. We establish convergence rate bounds of O(1 / t) and O(log t / t) for the synchronous and asynchronous cases respectively, where t is the number of iterations, with explicit data and network dependent terms. Beyond favorable comparisons in terms of rate analysis, numerical experiments provide empirical evidence the proposed algorithms surpasses the previously introduced approach.
****

Model-Based Relative Entropy Stochastic Search
Abbas Abdolmaleki, Rudolf Lioutikov, Jan R. Peters, Nuno Lau, Luis Pualo Reis, Gerhard Neumann
Stochastic search algorithms are general black-box optimizers. Due to their ease of use and their generality, they have recently also gained a lot of attention in operations research, machine learning and policy search. Yet, these algorithms require a lot of evaluations of the objective, scale poorly with the problem dimension, are affected by highly noisy objective functions and may converge prematurely. To alleviate these problems, we introduce a new surrogate-based stochastic search approach. We learn simple, quadratic surrogate models of the objective function. As the quality of such a quadratic approximation is limited, we do not greedily exploit the learned models. The algorithm can be misled by an inaccurate optimum introduced by the surrogate. Instead, we use information theoretic constraints to bound the `distance' between the new and old data distribution while maximizing the objective function. Additionally the new method is able to sustain the exploration of the search distribution to avoid premature convergence. We compare our method with state of art black-box optimization methods on standard uni-modal and multi-modal optimization functions, on simulated planar robot tasks and a complex robot ball throwing task.The proposed method considerably outperforms the existing approaches.
****

Semi-supervised Learning with Ladder Networks
Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, Tapani Raiko
We combine supervised learning with unsupervised learning in deep neural networks. The proposed model is trained to simultaneously minimize the sum of supervised and unsupervised cost functions by backpropagation, avoiding the need for layer-wise pre-training. Our work builds on top of the Ladder network proposed by Valpola (2015) which we extend by combining the model with supervision. We show that the resulting model reaches state-of-the-art performance in semi-supervised MNIST and CIFAR-10 classification in addition to permutation-invariant MNIST classification with all labels.
****

Empirical Localization of Homogeneous Divergences on Discrete Sample Spaces
Takashi Takenouchi, Takafumi Kanamori
In this paper, we propose a novel parameter estimator for probabilistic models on discrete space. The proposed estimator is derived from minimization of homogeneous divergence and can be constructed without calculation of the normalization constant, which is frequently infeasible for models in the discrete space. We in

vestigate statistical properties of the proposed estimator such as consistency and asymptotic normality, and reveal a relationship with the alpha-divergence. Small experiments show that the proposed estimator attains comparable performance to the MLE with drastically lower computational cost.

****

Enforcing balance allows local supervised learning in spiking recurrent networks

Ralph Bourdoukan, Sophie Denève

To predict sensory inputs or control motor trajectories, the brain must constantly learn temporal dynamics based on error feedback. However, it remains unclear how such supervised learning is implemented in biological neural networks. Learning in recurrent spiking networks is notoriously difficult because local changes in connectivity may have an unpredictable effect on the global dynamics. The most commonly used learning rules, such as temporal back-propagation, are not local and thus not biologically plausible. Furthermore, reproducing the Poisson-like statistics of neural responses requires the use of networks with balanced excitation and inhibition. Such balance is easily destroyed during learning. Using a top-down approach, we show how networks of integrate-and-fire neurons can learn arbitrary linear dynamical systems by feeding back their error as a feed-forward input. The network uses two types of recurrent connections: fast and slow. The fast connections learn to balance excitation and inhibition using a voltage-based plasticity rule. The slow connections are trained to minimize the error feedback using a current-based Hebbian learning rule. Importantly, the balance maintained by fast connections is crucial to ensure that global error signals are available locally in each neuron, in turn resulting in a local learning rule for the slow connections. This demonstrates that spiking networks can learn complex dynamics using purely local learning rules, using E/I balance as the key rather than an additional constraint. The resulting network implements a given function within the predictive coding scheme, with minimal dimensions and activity.

****

Online Learning for Adversaries with Memory: Price of Past Mistakes

Oren Anava, Elad Hazan, Shie Mannor

The framework of online learning with memory naturally captures learning problems with temporal effects, and was previously studied for the experts setting. In this work we extend the notion of learning with memory to the general Online Convex Optimization (OCO) framework, and present two algorithms that attain low regret. The first algorithm applies to Lipschitz continuous loss functions, obtaining optimal regret bounds for both convex and strongly convex losses. The second algorithm attains the optimal regret bounds and applies more broadly to convex losses without requiring Lipschitz continuity, yet is more complicated to implement. We complement the theoretic results with two applications: statistical arbitrage in finance, and multi-step ahead prediction in statistics.

****

Streaming, Distributed Variational Inference for Bayesian Nonparametrics

Trevor Campbell, Julian Straub, John W. Fisher III, Jonathan P. How

This paper presents a methodology for creating streaming, distributed inference algorithms for Bayesian nonparametric (BNP) models. In the proposed framework, processing nodes receive a sequence of data minibatches, compute a variational posterior for each, and make asynchronous streaming updates to a central model. In contrast to previous algorithms, the proposed framework is truly streaming, distributed, asynchronous, learning-rate-free, and truncation-free. The key challenge in developing the framework, arising from fact that BNP models do not impose an inherent ordering on their components, is finding the correspondence between minibatch and central BNP posterior components before performing each update. To address this, the paper develops a combinatorial optimization problem over component correspondences, and provides an efficient solution technique. The paper concludes with an application of the methodology to the DP mixture model, with experimental results demonstrating its practical scalability and performance.

****

Tree-Guided MCMC Inference for Normalized Random Measure Mixture Models

Juho Lee, Seungjin Choi

Normalized random measures (NRMs) provide a broad class of discrete random measures that are often used as priors for Bayesian nonparametric models. Dirichlet process is a well-known example of NRMs. Most of posterior inference methods for NRM mixture models rely on MCMC methods since they are easy to implement and their convergence is well studied. However, MCMC often suffers from slow convergence when the acceptance rate is low. Tree-based inference is an alternative deterministic posterior inference method, where Bayesian hierarchical clustering (BHC) or incremental Bayesian hierarchical clustering (IBHC) have been developed for DP or NRM mixture (NRMM) models, respectively. Although IBHC is a promising method for posterior inference for NRMM models due to its efficiency and applicability to online inference, its convergence is not guaranteed since it uses heuristics that simply selects the best solution after multiple trials are made. In this paper, we present a hybrid inference algorithm for NRMM models, which combines the merits of both MCMC and IBHC. Trees built by IBHC outlinespartitions of data, which guides Metropolis-Hastings procedure to employ appropriate proposals. Inheriting the nature of MCMC, our tree-guided MCMC (tgMCMC) is guaranteed to converge, and enjoys the fast convergence thanks to the effective proposals guided by trees. Experiments on both synthetic and real world datasets demonstrate the benefit of our method.

****

The Self-Normalized Estimator for Counterfactual Learning

Adith Swaminathan, Thorsten Joachims

This paper identifies a severe problem of the counterfactual risk estimator typically used in batch learning from logged bandit feedback (BLBF), and proposes the use of an alternative estimator that avoids this problem.In the BLBF setting, the learner does not receive full-information feedback like in supervised learning, but observes feedback only for the actions taken by a historical policy.This makes BLBF algorithms particularly attractive for training online systems (e.g., ad placement, web search, recommendation) using their historical logs.The Counterfactual Risk Minimization (CRM) principle offers a general recipe for designing BLBF algorithms. It requires a counterfactual risk estimator, and virtually all existing works on BLBF have focused on a particular unbiased estimator.We show that this conventional estimator suffers from apropensity overfitting problem when used for learning over complex hypothesis spaces.We propose to replace the risk estimator with a self-normalized estimator, showing that it neatly avoids this problem.This naturally gives rise to a new learning algorithm -- Normalized Policy Optimizer for Exponential Models (Norm-POEM) --for structured output prediction using linear rules.We evaluate the empirical effectiveness of Norm-POEM on severalmulti-label classification problems, finding that it consistently outperforms the conventional estimator.

****

Information-theoretic lower bounds for convex optimization with erroneous oracles

Yaron Singer, Jan Vondrak

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

A Nonconvex Optimization Framework for Low Rank Matrix Estimation

Tuo Zhao, Zhaoran Wang, Han Liu

We study the estimation of low rank matrices via nonconvex optimization. Compared with convex relaxation, nonconvex optimization exhibits superior empirical performance for large scale instances of low rank matrix estimation. However, the understanding of its theoretical guarantees are limited. In this paper, we define the notion of projected oracle divergence based on which we establish sufficient conditions for the success of nonconvex optimization. We illustrate the consequences of this general framework for matrix sensing and completion. In particular, we prove that a broad class of nonconvex optimization algorithms, including alternating minimization and gradient-type methods, geometrically converge to the global optimum and exactly recover the true low rank matrices under standard conditions.

****

Recursive Training of 2D-3D Convolutional Networks for Neuronal Boundary Prediction

Kisuk Lee, Aleksandar Zlateski, Vishwanathan Ashwin, H. Sebastian Seung

Efforts to automate the reconstruction of neural circuits from 3D electron microscopic (EM) brain images are critical for the field of connectomics. An important computation for reconstruction is the detection of neuronal boundaries. Images acquired by serial section EM, a leading 3D EM technique, are highly anisotropic, with inferior quality along the third dimension. For such images, the 2D max-pooling convolutional network has set the standard for performance at boundary detection. Here we achieve a substantial gain in accuracy through three innovations. Following the trend towards deeper networks for object recognition, we use a much deeper network than previously employed for boundary detection. Second, we incorporate 3D as well as 2D filters, to enable computations that use 3D context. Finally, we adopt a recursively trained architecture in which a first network generates a preliminary boundary map that is provided as input along with the original image to a second network that generates a final boundary map. Backpropagation training is accelerated by ZNN, a new implementation of 3D convolutional networks that uses multicore CPU parallelism for speed. Our hybrid 2D-3D architecture could be more generally applicable to other types of anisotropic 3D images, including video, and our recursive framework for any image labeling problem.

****

Multi-class SVMs: From Tighter Data-Dependent Generalization Bounds to Novel Algorithms

Yunwen Lei, Urun Dogan, Alexander Binder, Marius Kloft

This paper studies the generalization performance of multi-class classification algorithms, for which we obtain, for the first time, a data-dependent generalization error bound with a logarithmic dependence on the class size, substantially improving the state-of-the-art linear dependence in the existing data-dependent generalization analysis. The theoretical analysis motivates us to introduce a new multi-class classification machine based on lp-norm regularization, where the parameter p controls the complexity of the corresponding bounds. We derive an efficient optimization algorithm based on Fenchel duality theory. Benchmarks on several real-world datasets show that the proposed algorithm can achieve significant accuracy gains over the state of the art.

****

Scalable Inference for Gaussian Process Models with Black-Box Likelihoods

Amir Dezfouli, Edwin V. Bonilla

We propose a sparse method for scalable automated variational inference (AVI) in a large class of models with Gaussian process (GP) priors, multiple latent functions, multiple outputs and non-linear likelihoods. Our approach maintains the statistical efficiency property of the original AVI method, requiring only expectations over univariate Gaussian distributions to approximate the posterior with a mixture of Gaussians. Experiments on small datasets for various problems including regression, classification, Log Gaussian Cox processes, and warped GPs show that our method can perform as well as the full method under high levels of sparsity. On larger experiments using the MNIST and the SARCOS datasets we show that our method can provide superior performance to previously published scalable approaches that have been handcrafted to specific likelihood models.

****

M-Best-Diverse Labelings for Submodular Energies and Beyond

Alexander Kirillov, Dmytro Shlezinger, Dmitry P. Vetrov, Carsten Rother, Bogdan Savchynskyy

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

BinaryConnect: Training Deep Neural Networks with binary weights during propagations

Matthieu Courbariaux, Yoshua Bengio, Jean-Pierre David

Deep Neural Networks (DNN) have achieved state-of-the-art results in a wide range of tasks, with the best results obtained with large training sets and large models. In the past, GPUs enabled these breakthroughs because of their greater computational speed. In the future, faster computation at both training and test time is likely to be crucial for further progress and for consumer applications on low-power devices. As a result, there is much interest in research and development of dedicated hardware for Deep Learning (DL). Binary weights, i.e., weights which are constrained to only two possible values (e.g. -1 or 1), would bring great benefits to specialized DL hardware by replacing many multiply-accumulate operations by simple accumulations, as multipliers are the most space and power-hungry components of the digital implementation of neural networks. We introduce BinaryConnect, a method which consists in training a DNN with binary weights during the forward and backward propagations, while retaining precision of the stored weights in which gradients are accumulated. Like other dropout schemes, we show that BinaryConnect acts as regularizer and we obtain near state-of-the-art results with BinaryConnect on the permutation-invariant MNIST, CIFAR-10 and SVHN.

****

No-Regret Learning in Bayesian Games

Jason Hartline, Vasilis Syrgkanis, Eva Tardos

Recent price-of-anarchy analyses of games of complete information suggest that coarse correlated equilibria, which characterize outcomes resulting from no-regret learning dynamics, have near-optimal welfare. This work provides two main technical results that lift this conclusion to games of incomplete information, a.k.a., Bayesian games. First, near-optimal welfare in Bayesian games follows directly from the smoothness-based proof of near-optimal welfare in the same game when the private information is public. Second, no-regret learning dynamics converge to Bayesian coarse correlated equilibrium in these incomplete information games. These results are enabled by interpretation of a Bayesian game as a stochastic game of complete information.

****

Robust Gaussian Graphical Modeling with the Trimmed Graphical Lasso

Eunho Yang, Aurelie C. Lozano

Gaussian Graphical Models (GGMs) are popular tools for studying network structures. However, many modern applications such as gene network discovery and social interactions analysis often involve high-dimensional noisy data with outliers or heavier tails than the Gaussian distribution. In this paper, we propose the Trimmed Graphical Lasso for robust estimation of sparse GGMs. Our method guards against outliers by an implicit trimming mechanism akin to the popular Least Trimmed Squares method used for linear regression. We provide a rigorous statistical analysis of our estimator in the high-dimensional setting. In contrast, existing approaches for robust sparse GGMs estimation lack statistical guarantees. Our theoretical results are complemented by experiments on simulated and real gene expression data which further demonstrate the value of our approach.

****

Parallelizing MCMC with Random Partition Trees

Xiangyu Wang, Fangjian Guo, Katherine A. Heller, David B. Dunson

The modern scale of data has brought new challenges to Bayesian inference. In particular, conventional MCMC algorithms are computationally very expensive for large data sets. A promising approach to solve this problem is embarrassingly parallel MCMC (EP-MCMC), which first partitions the data into multiple subsets and runs independent sampling algorithms on each subset. The subset posterior draws are then aggregated via some combining rules to obtain the final approximation. Existing EP-MCMC algorithms are limited by approximation accuracy and difficulty in resampling. In this article, we propose a new EP-MCMC algorithm PART that solves these problems. The new algorithm applies random partition trees to combine the subset posterior draws, which is distribution-free, easy to resample from and can adapt to multiple scales. We provide theoretical justification and extens ive experiments illustrating empirical performance.

****

Convergence rates of sub-sampled Newton methods

Murat A. Erdogdu, Andrea Montanari

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Learning Theory and Algorithms for Forecasting Non-stationary Time Series

Vitaly Kuznetsov, Mehryar Mohri

We present data-dependent learning bounds for the general scenario of non-stationary non-mixing stochastic processes. Our learning guarantees are expressed in terms of a data-dependent measure of sequential complexity and a discrepancy measure that can be estimated from data under some mild assumptions. We use our learning bounds to devise new algorithms for non-stationary time series forecasting for which we report some preliminary experimental results.

****

Equilibrated adaptive learning rates for non-convex optimization

Yann Dauphin, Harm de Vries, Yoshua Bengio

Parameter-specific adaptive learning rate methods are computationally efficient ways to reduce the ill-conditioning problems encountered when training large deep networks. Following recent work that strongly suggests that most of thecritical points encountered when training such networks are saddle points, we find how considering the presence of negative eigenvalues of the Hessian could help us design better suited adaptive learning rate schemes. We show that the popular Jacobi preconditioner has undesirable behavior in the presence of both positive and negative curvature, and present theoretical and empirical evidence that the so-called equilibration preconditioner is comparatively better suited to non-convex problems. We introduce a novel adaptive learning rate scheme, called ESGD, based on the equilibration preconditioner. Our experiments demonstrate that both schemes yield very similar step directions but that ESGD sometimes surpasses RMSProp in terms of convergence speed, always clearly improving over plain stochastic gradient descent.

****

Optimal Linear Estimation under Unknown Nonlinear Transform

Xinyang Yi, Zhaoran Wang, Constantine Caramanis, Han Liu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Analysis of Robust PCA via Local Incoherence

Huishuai Zhang, Yi Zhou, Yingbin Liang

We investigate the robust PCA problem of decomposing an observed matrix into the sum of a low-rank and a sparse error matrices via convex programming Principal Component Pursuit (PCP). In contrast to previous studies that assume the support of the error matrix is generated by uniform Bernoulli sampling, we allow non-uniform sampling, i.e., entries of the low-rank matrix are corrupted by errors with unequal probabilities. We characterize conditions on error corruption of each individual entry based on the local incoherence of the low-rank matrix, under which correct matrix decomposition by PCP is guaranteed. Such a refined analysis of robust PCA captures how robust each entry of the low rank matrix combats error corruption. In order to deal with non-uniform error corruption, our technical proof introduces a new weighted norm and develops/exploits the concentration properties that such a norm satisfies.

****

Probabilistic Variational Bounds for Graphical Models

Qiang Liu, John W. Fisher III, Alexander T. Ihler

Variational algorithms such as tree-reweighted belief propagation can provide de terministic bounds on the partition function, but are often loose and difficult to use in an any-time'' fashion, expending more computation for tighter bounds. On the other hand, Monte Carlo estimators such as importance sampling have excellent any-time behavior, but depend critically on the proposal distribution. We propose a simple Monte Carlo based inference method that augments convex variational bounds by adding importance sampling (IS). We argue that convex variational methods naturally provide good IS proposals thatcover the probability of the target distribution, and reinterpret the variational optimization as designing a proposal to minimizes an upper bound on the variance of our IS estimator. This both provides an accurate estimator and enables the construction of any-time probabilistic bounds that improve quickly and directly on state of-the-art variational bounds, which provide certificates of accuracy given enough samples relative to the error in the initial bound.

****

The Human Kernel

Andrew G. Wilson, Christoph Dann, Chris Lucas, Eric P. Xing

Bayesian nonparametric models, such as Gaussian processes, provide a compelling framework for automatic statistical modelling: these models have a high degree of flexibility, and automatically calibrated complexity. However, automating human expertise remains elusive; for example, Gaussian processes with standard kernels struggle on function extrapolation problems that are trivial for human learners. In this paper, we create function extrapolation problems and acquire human responses, and then design a kernel learning framework to reverse engineer the inductive biases of human learners across a set of behavioral experiments. We use the learned kernels to gain psychological insights and to extrapolate in human-like ways that go beyond traditional stationary and polynomial kernels. Finally, we investigate Occam's razor in human and Gaussian process based function learning.

****

Asynchronous Parallel Stochastic Gradient for Nonconvex Optimization

Xiangru Lian, Yijun Huang, Yuncheng Li, Ji Liu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Evaluating the statistical significance of biclusters

Jason D. Lee, Yuekai Sun, Jonathan E. Taylor

Biclustering (also known as submatrix localization) is a problem of high practical relevance in exploratory analysis of high-dimensional data. We develop a framework for performing statistical inference on biclusters found by score-based algorithms. Since the bicluster was selected in a data dependent manner by a biclustering or localization algorithm, this is a form of selective inference. Our framework gives exact (non-asymptotic) confidence intervals and p-values for the significance of the selected biclusters. Further, we generalize our approach to obtain exact inference for Gaussian statistics.

****

Fast and Guaranteed Tensor Decomposition via Sketching

Yining Wang, Hsiao-Yu Tung, Alexander J. Smola, Anima Anandkumar

Tensor CANDECOMP/PARAFAC (CP) decomposition has wide applications in statistical learning of latent variable models and in data mining. In this paper, we propose fast and randomized tensor CP decomposition algorithms based on sketching. We build on the idea of count sketches, but introduce many novel ideas which are unique to tensors. We develop novel methods for randomized com- putation of tensor contractions via FFTs, without explicitly forming the tensors. Such tensor contractions are encountered in decomposition methods such as ten- sor power iterations and alternating least squares. We also design novel colliding hashes for symmetric tensors to further save time in computing the sketches. We then combine these sketching ideas with existing whitening and tensor power iter- ative techniques to obtain the fastest algorithm on both sparse and dense tensors. The quali ty of approximation under our method does not depend on properties such as sparsity, uniformity of elements, etc. We apply the method for topic mod- eling and obtain competitive results.

****

Inverse Reinforcement Learning with Locally Consistent Reward Functions

Quoc Phong Nguyen, Bryan Kian Hsiang Low, Patrick Jaillet

Existing inverse reinforcement learning (IRL) algorithms have assumed each expert’s demonstrated trajectory to be produced by only a single reward function. This paper presents a novel generalization of the IRL problem that allows each trajectory to be generated by multiple locally consistent reward functions, hence catering to more realistic and complex experts’ behaviors. Solving our generalized IRL problem thus involves not only learning these reward functions but also the stochastic transitions between them at any state (including unvisited states). By representing our IRL problem with a probabilistic graphical model, an expectation-maximization (EM) algorithm can be devised to iteratively learn the different reward functions and the stochastic transitions between them in order to jointly improve the likelihood of the expert’s demonstrated trajectories. As a result, the most likely partition of a trajectory into segments that are generated from different locally consistent reward functions selected by EM can be derived. Empirical evaluation on synthetic and real-world datasets shows that our IRL algorithm outperforms the state-of-the-art EM clustering with maximum likelihood IRL, which is, interestingly, a reduced variant of our approach.

****

A hybrid sampler for Poisson-Kingman mixture models

Maria Lomeli, Stefano Favaro, Yee Whye Teh

This paper concerns the introduction of a new Markov Chain Monte Carlo scheme for posterior sampling in Bayesian nonparametric mixture models with priors that belong to the general Poisson-Kingman class. We present a novel and compact way of representing the infinite dimensional component of the model such that while explicitly representing this infinite component it has less memory and storage requirements than previous MCMC schemes. We describe comparative simulation results demonstrating the efficacy of the proposed MCMC algorithm against existing marginal and conditional MCMC samplers.

****

Learning with Symmetric Label Noise: The Importance of Being Unhinged

Brendan van Rooyen, Aditya Menon, Robert C. Williamson

Convex potential minimisation is the de facto approach to binary classification. However, Long and Servedio [2008] proved that under symmetric label noise (SLN), minimisation of any convex potential over a linear function class can result in classification performance equivalent to random guessing. This ostensibly shows that convex losses are not SLN-robust. In this paper, we propose a convex, classification-calibrated loss and prove that it is SLN-robust. The loss avoids the Long and Servedio [2008] result by virtue of being negatively unbounded. The loss is a modification of the hinge loss, where one does not clamp at zero; hence, we call it the unhinged loss. We show that the optimal unhinged solution is equivalent to that of a strongly regularised SVM, and is the limiting solution for any convex potential; this implies that strong l2 regularisation makes most standard learners SLN-robust. Experiments confirm the unhinged loss’ SLN-robustness.

****

Visalogy: Answering Visual Analogy Questions

Fereshteh Sadeghi, C. Lawrence Zitnick, Ali Farhadi

In this paper, we study the problem of answering visual analogy questions. These questions take the form of image A is to image B as image C is to what. Answering these questions entails discovering the mapping from image A to image B and then extending the mapping to image C and searching for the image D such that the relation from A to B holds for C to D. We pose this problem as learning an embedding that encourages pairs of analogous images with similar transformations to be close together using convolutional neural networks with a quadruple Siamese architecture. We introduce a dataset of visual analogy questions in natural images, and show first results of its kind on solving analogy questions on natural im ages.

****

Cornering Stationary and Restless Mixing Bandits with Remix-UCB

Julien Audiffren, Liva Ralaivola

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

The Consistency of Common Neighbors for Link Prediction in Stochastic Blockmodels

Purnamrita Sarkar, Deepayan Chakrabarti, peter j. bickel

Link prediction and clustering are key problems for network-structureddata. While spectral clustering has strong theoretical guaranteesunder the popular stochastic blockmodel formulation of networks, itcan be expensive for large graphs. On the other hand, the heuristic ofpredicting links to nodes that share the most common neighbors withthe query node is much fast, and works very well in practice. We showtheoretically that the common neighbors heuristic can extract clustersw.h.p. when the graph is dense enough, and can do so even in sparsergraphs with the addition of a ``cleaning'' step. Empirical results onsimulated and real-world data support our conclusions.

****

On the Accuracy of Self-Normalized Log-Linear Models

Jacob Andreas, Maxim Rabinovich, Michael I. Jordan, Dan Klein

Calculation of the log-normalizer is a major computational obstacle in applications of log-linear models with large output spaces. The problem of fast normalizer computation has therefore attracted significant attention in the theoretical and applied machine learning literature. In this paper, we analyze a recently proposed technique known as ``self-normalization'', which introduces a regularization term in training to penalize log normalizers for deviating from zero. This makes it possible to use unnormalized model scores as approximate probabilities. Empirical evidence suggests that self-normalization is extremely effective, but a theoretical understanding of why it should work, and how generally it can be applied, is largely lacking.We prove upper bounds on the loss in accuracy due to self-normalization, describe classes of input distributionsthat self-normalize easily, and construct explicit examples of high-variance input distributions. Our theoretical results make predictions about the difficulty of fitting self-normalized models to several classes of distributions, and we conclude with empirical validation of these predictions on both real and synthetic datasets.

****

Learnability of Influence in Networks

Harikrishna Narasimhan, David C. Parkes, Yaron Singer

We establish PAC learnability of influence functions for three common influence models, namely, the Linear Threshold (LT), Independent Cascade (IC) and Voter models, and present concrete sample complexity results in each case. Our results for the LT model are based on interesting connections with neural networks; those for the IC model are based an interpretation of the influence function as an expectation over random draw of a subgraph and use covering number arguments; and those for the Voter model are based on a reduction to linear regression. We show these results for the case in which the cascades are only partially observed and we do not see the time steps in which a node has been influenced. We also provide efficient polynomial time learning algorithms for a setting with full observation, i.e. where the cascades also contain the time steps in which nodes are influenced.

****

Linear Response Methods for Accurate Covariance Estimates from Mean Field Variational Bayes

Ryan J. Giordano, Tamara Broderick, Michael I. Jordan

Mean field variational Bayes (MFVB) is a popular posterior approximation method due to its fast runtime on large-scale data sets. However, a well known failing

of MFVB is that it underestimates the uncertainty of model variables (sometimes severely) and provides no information about model variable covariance. We generalize linear response methods from statistical physics to deliver accurate uncertainty estimates for model variables---both for individual variables and coherently across variables. We call our method linear response variational Bayes (LRVB). When the MFVB posterior approximation is in the exponential family, LRVB has a simple, analytic form, even for non-conjugate models. Indeed, we make no assumptions about the form of the true posterior. We demonstrate the accuracy and scalability of our method on a range of models for both simulated and real data.
****

Weighted Theta Functions and Embeddings with Applications to Max-Cut, Clustering and Summarization

Fredrik D. Johansson, Ankani Chattoraj, Chiranjib Bhattacharyya, Devdatt Dubhashi

We introduce a unifying generalization of the Lovász theta function, and the associated geometric embedding, for graphs with weights on both nodes and edges. We show how it can be computed exactly by semidefinite programming, and how to approximate it using SVM computations. We show how the theta function can be interpreted as a measure of diversity in graphs and use this idea, and the graph embedding in algorithms for Max-Cut, correlation clustering and document summarization, all of which are well represented as problems on weighted graphs.
****

End-to-end Learning of LDA by Mirror-Descent Back Propagation over a Deep Architecture

Jianshu Chen, Ji He, Yelong Shen, Lin Xiao, Xiaodong He, Jianfeng Gao, Xinying Song, Li Deng

We develop a fully discriminative learning approach for supervised Latent Dirichlet Allocation (LDA) model using Back Propagation (i.e., BP-sLDA), which maximizes the posterior probability of the prediction variable given the input document. Different from traditional variational learning or Gibbs sampling approaches, the proposed learning method applies (i) the mirror descent algorithm for maximum a posterior inference and (ii) back propagation over a deep architecture together with stochastic gradient/mirror descent for model parameter estimation, leading to scalable and end-to-end discriminative learning of the model. As a byproduct, we also apply this technique to develop a new learning method for the traditional unsupervised LDA model (i.e., BP-LDA). Experimental results on three real-world regression and classification tasks show that the proposed methods significantly outperform the previous supervised topic models, neural networks, and is on par with deep neural networks.
****

Robust Spectral Inference for Joint Stochastic Matrix Factorization

Moontae Lee, David Bindel, David Mimno

Spectral inference provides fast algorithms and provable optimality for latent topic analysis. But for real data these algorithms require additional ad-hoc heuristics, and even then often produce unusable results. We explain this poor performance by casting the problem of topic inference in the framework of Joint Stochastic Matrix Factorization (JSMF) and showing that previous methods violate the theoretical conditions necessary for a good solution to exist. We then propose a novel rectification method that learns high quality topics and their interactions even on small, noisy data. This method achieves results comparable to probabilistic techniques in several domains while maintaining scalability and provable optimality.
****

Minimax Time Series Prediction

Wouter M. Koolen, Alan Malek, Peter L. Bartlett, Yasin Abbasi Yadkori

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Learning to Segment Object Candidates
Pedro O. O. Pinheiro, Ronan Collobert, Piotr Dollar

Recent object detection systems rely on two critical steps: (1) a set of object proposals is predicted as efficiently as possible, and (2) this set of candidate proposals is then passed to an object classifier. Such approaches have been shown they can be fast, while achieving the state of the art in detection performance. In this paper, we propose a new way to generate object proposals, introducing an approach based on a discriminative convolutional network. Our model is trained jointly with two objectives: given an image patch, the first part of the system outputs a class-agnostic segmentation mask, while the second part of the system outputs the likelihood of the patch being centered on a full object. At test time, the model is efficiently applied on the whole test image and generates a set of segmentation masks, each of them being assigned with a corresponding object likelihood score. We show that our model yields significant improvements over state-of-the-art object proposal algorithms. In particular, compared to previous approaches, our model obtains substantially higher object recall using fewer proposals. We also show that our model is able to generalize to unseen categories it has not seen during training. Unlike all previous approaches for generating object masks, we do not rely on edges, superpixels, or any other form of low-level segmentation.

****

A Theory of Decision Making Under Dynamic Context
Michael Shvartsman, Vaibhav Srivastava, Jonathan D. Cohen

The dynamics of simple decisions are well understood and modeled as a class of random walk models (e.g. Laming, 1968; Ratcliff, 1978; Busemeyer and Townsend, 1993; Usher and McClelland, 2001; Bogacz et al., 2006). However, most real-life decisions include a rich and dynamically-changing influence of additional information we call context. In this work, we describe a computational theory of decision making under dynamically shifting context. We show how the model generalizes the dominant existing model of fixed-context decision making (Ratcliff, 1978) and can be built up from a weighted combination of fixed-context decisions evolving simultaneously. We also show how the model generalizes re- cent work on the control of attention in the Flanker task (Yu et al., 2009). Finally, we show how the model recovers qualitative data patterns in another task of longstanding psychological interest, the AX Continuous Performance Test (Servan-Schreiber et al., 1996), using the same model parameters.

****

Particle Gibbs for Infinite Hidden Markov Models
Nilesh Tripuraneni, Shixiang (Shane) Gu, Hong Ge, Zoubin Ghahramani

Infinite Hidden Markov Models (iHMM's) are an attractive, nonparametric generalization of the classical Hidden Markov Model which can automatically infer the number of hidden states in the system. However, due to the infinite-dimensional nature of the transition dynamics, performing inference in the iHMM is difficult. In this paper, we present an infinite-state Particle Gibbs (PG) algorithm to resample state trajectories for the iHMM. The proposed algorithm uses an efficient proposal optimized for iHMMs, and leverages ancestor sampling to improve the mixing of the standard PG algorithm. Our algorithm demonstrates significant convergence improvements on synthetic and real world data sets.

****

Bandit Smooth Convex Optimization: Improving the Bias-Variance Tradeoff
Ofer Dekel, Ronen Eldan, Tomer Koren

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Compressive spectral embedding: sidestepping the SVD
Dinesh Ramasamy, Upamanyu Madhow

Spectral embedding based on the Singular Value Decomposition (SVD) is a widely used preprocessing step in many learning tasks, typically leading to dimensionali ty reduction by projecting onto a number of dominant singular vectors and rescaling the coordinate axes (by a predefined function of the singular value). However, the number of such vectors required to capture problem structure grows with problem size, and even partial SVD computation becomes a bottleneck. In this paper, we propose a low-complexity it compressive spectral embedding algorithm, which employs random projections and finite order polynomial expansions to compute approximations to SVD-based embedding. For an m times n matrix with T non-zeros, its time complexity is O((T+m+n)log(m+n)), and the embedding dimension is O(log(m+n)), both of which are independent of the number of singular vectors whose effect we wish to capture. To the best of our knowledge, this is the first work to circumvent this dependence on the number of singular vectors for general SVD-based embeddings. The key to sidestepping the SVD is the observation that, for downstream inference tasks such as clustering and classification, we are only interested in using the resulting embedding to evaluate pairwise similarity metrics derived from the euclidean norm, rather than capturing the effect of the underlying matrix on arbitrary vectors as a partial SVD tries to do. Our numerical results on network datasets demonstrate the efficacy of the proposed method, and motivate further exploration of its application to large-scale inference tasks.

****

Winner-Take-All Autoencoders

Alireza Makhzani, Brendan J. Frey

In this paper, we propose a winner-take-all method for learning hierarchical sparse representations in an unsupervised fashion. We first introduce fully-connected winner-take-all autoencoders which use mini-batch statistics to directly enforce a lifetime sparsity in the activations of the hidden units. We then propose the convolutional winner-take-all autoencoder which combines the benefits of convolutional architectures and autoencoders for learning shift-invariant sparse representations. We describe a way to train convolutional autoencoders layer by layer, where in addition to lifetime sparsity, a spatial sparsity within each feature map is achieved using winner-take-all activation functions. We will show that winner-take-all autoencoders can be used to to learn deep sparse representations from the MNIST, CIFAR-10, ImageNet, Street View House Numbers and Toronto Face datasets, and achieve competitive classification performance.

****

Robust Feature-Sample Linear Discriminant Analysis for Brain Disorders Diagnosis

Ehsan Adeli-Mosabbeb, Kim-Han Thung, Le An, Feng Shi, Dinggang Shen

A wide spectrum of discriminative methods is increasingly used in diverse applications for classification or regression tasks. However, many existing discriminative methods assume that the input data is nearly noise-free, which limits their applications to solve real-world problems. Particularly for disease diagnosis, the data acquired by the neuroimaging devices are always prone to different sources of noise. Robust discriminative models are somewhat scarce and only a few attempts have been made to make them robust against noise or outliers. These methods focus on detecting either the sample-outliers or feature-noises. Moreover, they usually use unsupervised de-noising procedures, or separately de-noise the training and the testing data. All these factors may induce biases in the learning process, and thus limit its performance. In this paper, we propose a classification method based on the least-squares formulation of linear discriminant analysis, which simultaneously detects the sample-outliers and feature-noises. The proposed method operates under a semi-supervised setting, in which both labeled training and unlabeled testing data are incorporated to form the intrinsic geometry of the sample space. Therefore, the violating samples or feature values are identified as sample-outliers or feature-noises, respectively. We test our algorithm on one synthetic and two brain neurodegenerative databases (particularly for Parkinson's disease and Alzheimer's disease). The results demonstrate that our method outperforms all baseline and state-of-the-art methods, in terms of both accuracy and the area under the ROC curve.

****

COEVOLVE: A Joint Point Process Model for Information Diffusion and Network Co-evolution

Mehrdad Farajtabar, Yichen Wang, Manuel Gomez Rodriguez, Shuang Li, Hongyuan Zha, Le Song

Information diffusion in online social networks is affected by the underlying network topology, but it also has the power to change it. Online users are constantly creating new links when exposed to new information sources, and in turn these links are alternating the way information spreads. However, these two highly intertwined stochastic processes, information diffusion and network evolution, have been predominantly studied separately, ignoring their co-evolutionary dynamics.We propose a temporal point process model, COEVOLVE, for such joint dynamics, allowing the intensity of one process to be modulated by that of the other. This model allows us to efficiently simulate interleaved diffusion and network events, and generate traces obeying common diffusion and network patterns observed in real-world networks. Furthermore, we also develop a convex optimization framework to learn the parameters of the model from historical diffusion and network evolution traces. We experimented with both synthetic data and data gathered from Twitter, and show that our model provides a good fit to the data as well as more accurate predictions than alternatives.

****

Nearly Optimal Private LASSO

Kunal Talwar, Abhradeep Guha Thakurta, Li Zhang

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Calibrated Structured Prediction

Volodymyr Kuleshov, Percy S. Liang

In user-facing applications, displaying calibrated confidence measures---probabilities that correspond to true frequency---can be as important as obtaining high accuracy. We are interested in calibration for structured prediction problems such as speech recognition, optical character recognition, and medical diagnosis. Structured prediction presents new challenges for calibration: the output space is large, and users may issue many types of probability queries (e.g., marginals) on the structured output. We extend the notion of calibration so as to handle various subtleties pertaining to the structured setting, and then provide a simple recalibration method that trains a binary classifier to predict probabilities of interest. We explore a range of features appropriate for structured recalibration, and demonstrate their efficacy on three real-world datasets.

****

Spectral Representations for Convolutional Neural Networks

Oren Rippel, Jasper Snoek, Ryan P. Adams

Discrete Fourier transforms provide a significant speedup in the computation of convolutions in deep learning. In this work, we demonstrate that, beyond its advantages for efficient computation, the spectral domain also provides a powerful representation in which to model and train convolutional neural networks (CNNs). We employ spectral representations to introduce a number of innovations to CNN design. First, we propose spectral pooling, which performs dimensionality reduction by truncating the representation in the frequency domain. This approach preserves considerably more information per parameter than other pooling strategies and enables flexibility in the choice of pooling output dimensionality. This representation also enables a new form of stochastic regularization by randomized modification of resolution. We show that these methods achieve competitive results on classification and approximation tasks, without using any dropout or max-pooling. Finally, we demonstrate the effectiveness of complex-coefficient spectral parameterization of convolutional filters. While this leaves the underlying model unchanged, it results in a representation that greatly facilitates optimization. We observe on a variety of popular CNN configurations that this leads to significantly faster convergence during training.

****

On the consistency theory of high dimensional variable screening

Xiangyu Wang, Chenlei Leng, David B. Dunson

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Revenue Optimization against Strategic Buyers

Mehryar Mohri, Andres Munoz

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

The Population Posterior and Bayesian Modeling on Streams

James McInerney, Rajesh Ranganath, David Blei

Many modern data analysis problems involve inferences from streaming data. However, streaming data is not easily amenable to the standard probabilistic modeling approaches, which assume that we condition on finite data. We develop population variational Bayes, a new approach for using Bayesian modeling to analyze streams of data. It approximates a new type of distribution, the population posterior, which combines the notion of a population distribution of the data with Bayesian inference in a probabilistic model. We study our method with latent Dirichlet allocation and Dirichlet process mixtures on several large-scale data sets.

****

Parallel Predictive Entropy Search for Batch Global Optimization of Expensive Objective Functions

Amar Shah, Zoubin Ghahramani

We develop \textit{parallel predictive entropy search} (PPES), a novel algorithm for Bayesian optimization of expensive black-box objective functions. At each iteration, PPES aims to select a \textit{batch} of points which will maximize the information gain about the global maximizer of the objective. Well known strategies exist for suggesting a single evaluation point based on previous observations, while far fewer are known for selecting batches of points to evaluate in parallel. The few batch selection schemes that have been studied all resort to greedy methods to compute an optimal batch. To the best of our knowledge, PPES is the first non-greedy batch Bayesian optimization strategy. We demonstrate the benefit of this approach in optimization performance on both synthetic and real world applications, including problems in machine learning, rocket science and robotics.

****

The Return of the Gating Network: Combining Generative Models and Discriminative Training in Natural Image Priors

Dan Rosenbaum, Yair Weiss

In recent years, approaches based on machine learning have achieved state-of-the-art performance on image restoration problems. Successful approaches include both generative models of natural images as well as discriminative training of deep neural networks. Discriminative training of feed forward architectures allows explicit control over the computational cost of performing restoration and therefore often leads to better performance at the same cost at run time. In contrast, generative models have the advantage that they can be trained once and then adapted to any image restoration task by a simple use of Bayes' rule. In this paper we show how to combine the strengths of both approaches by training a discriminative, feed-forward architecture to predict the state of latent variables in a generative model of natural images. We apply this idea to the very successful Gaussian Mixture Model (GMM) of natural images. We show that it is possible to achieve comparable performance as the original GMM but with two orders of magnitude improvement in run time while maintaining the advantage of generative models.

****

Fighting Bandits with a New Kind of Smoothness

Jacob D. Abernethy, Chansoo Lee, Ambuj Tewari

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Sparse and Low-Rank Tensor Decomposition

Parikshit Shah, Nikhil Rao, Gongguo Tang

Motivated by the problem of robust factorization of a low-rank tensor, we study the question of sparse and low-rank tensor decomposition. We present an efficient computational algorithm that modifies Leurgans' algoirthm for tensor factorization. Our method relies on a reduction of the problem to sparse and low-rank matrix decomposition via the notion of tensor contraction. We use well-understood convex techniques for solving the reduced matrix sub-problem which then allows us to perform the full decomposition of the tensor. We delineate situations where the problem is recoverable and provide theoretical guarantees for our algorithm. We validate our algorithm with numerical experiments.

****

Testing Closeness With Unequal Sized Samples

Bhaswar Bhattacharya, Gregory Valiant

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach

Yinlam Chow, Aviv Tamar, Shie Mannor, Marco Pavone

In this paper we address the problem of decision making within a Markov decision process (MDP) framework where risk and modeling errors are taken into account. Our approach is to minimize a risk-sensitive conditional-value-at-risk (CVaR) objective, as opposed to a standard risk-neutral expectation. We refer to such problem as CVaR MDP. Our first contribution is to show that a CVaR objective, besides capturing risk sensitivity, has an alternative interpretation as expected cost under worst-case modeling errors, for a given error budget. This result, which is of independent interest, motivates CVaR MDPs as a unifying framework for risk-sensitive and robust decision making. Our second contribution is to present a value-iteration algorithm for CVaR MDPs, and analyze its convergence rate. To our knowledge, this is the first solution algorithm for CVaR MDPs that enjoys error guarantees. Finally, we present results from numerical experiments that corroborate our theoretical findings and show the practicality of our approach.

****

Fast Lifted MAP Inference via Partitioning

Somdeb Sarkhel, Parag Singla, Vibhav G. Gogate

Recently, there has been growing interest in lifting MAP inference algorithms for Markov logic networks (MLNs). A key advantage of these lifted algorithms is that they have much smaller computational complexity than propositional algorithms when symmetries are present in the MLN and these symmetries can be detected using lifted inference rules. Unfortunately, lifted inference rules are sound but not complete and can often miss many symmetries. This is problematic because when symmetries cannot be exploited, lifted inference algorithms ground the MLN, and search for solutions in the much larger propositional space. In this paper, we present a novel approach, which cleverly introduces new symmetries at the time of grounding. Our main idea is to partition the ground atoms and force the inference algorithm to treat all atoms in each part as indistinguishable. We show that by systematically and carefully refining (and growing) the partitions, we can build advanced any-time and any-space MAP inference algorithms. Our experiments on several real-world datasets clearly show that our new algorithm is superior to previous approaches and often finds useful symmetries in the search space that existing lifted inference rules are unable to detect.

****

Algorithmic Stability and Uniform Generalization

Ibrahim M. Alabdulmohsin

One of the central questions in statistical learning theory is to determine the conditions under which agents can learn from experience. This includes the necessary and sufficient conditions for generalization from a given finite training set to new observations. In this paper, we prove that algorithmic stability in the inference process is equivalent to uniform generalization across all parametric loss functions. We provide various interpretations of this result. For instance, a relationship is proved between stability and data processing, which reveals that algorithmic stability can be improved by post-processing the inferred hypothesis or by augmenting training examples with artificial noise prior to learning. In addition, we establish a relationship between algorithmic stability and the size of the observation space, which provides a formal justification for dimensionality reduction methods. Finally, we connect algorithmic stability to the size of the hypothesis space, which recovers the classical PAC result that the size (complexity) of the hypothesis space should be controlled in order to improve algorithmic stability and improve generalization.

****

Learning with Group Invariant Features: A Kernel Perspective.

Youssef Mroueh, Stephen Voinea, Tomaso A. Poggio

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Tractable Bayesian Network Structure Learning with Bounded Vertex Cover Number

Janne H. Korhonen, Pekka Parviainen

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Convergence Analysis of Prediction Markets via Randomized Subspace Descent

Rafael Frongillo, Mark D. Reid

Prediction markets are economic mechanisms for aggregating information about future events through sequential interactions with traders. The pricing mechanisms in these markets are known to be related to optimization algorithms in machine learning and through these connections we have some understanding of how equilibrium market prices relate to the beliefs of the traders in a market. However, little is known about rates and guarantees for the convergence of these sequential mechanisms, and two recent papers cite this as an important open question.In this paper we show how some previously studied prediction market trading models can be understood as a natural generalization of randomized coordinate descent which we call randomized subspace descent (RSD). We establish convergence rates for RSD and leverage them to prove rates for the two prediction market models above, answering the open questions. Our results extend beyond standard centralized markets to arbitrary trade networks.

****

SGD Algorithms based on Incomplete U-statistics: Large-Scale Minimization of Empirical Risk

Guillaume Papa, Stéphan Clémençon, Aurélien Bellet

In many learning problems, ranging from clustering to ranking through metric learning, empirical estimates of the risk functional consist of an average over tuples (e.g., pairs or triplets) of observations, rather than over individual observations. In this paper, we focus on how to best implement a stochastic approximation approach to solve such risk minimization problems. We argue that in the large-scale setting, gradient estimates should be obtained by sampling tuples of data points with replacement (incomplete U-statistics) instead of sampling data points without replacement (complete U-statistics based on subsamples). We develop a theoretical framework accounting for the substantial impact of this strategy on the generalization ability of the prediction model returned by the Stochastic

Gradient Descent (SGD) algorithm. It reveals that the method we promote achieves a much better trade-off between statistical accuracy and computational cost. Beyond the rate bound analysis, experiments on AUC maximization and metric learning provide strong empirical evidence of the superiority of the proposed approach.

****

Multi-Layer Feature Reduction for Tree Structured Group Lasso via Hierarchical Projection

Jie Wang, Jieping Ye

Tree structured group Lasso (TGL) is a powerful technique in uncovering the tree structured sparsity over the features, where each node encodes a group of features. It has been applied successfully in many real-world applications. However, with extremely large feature dimensions, solving TGL remains a significant challenge due to its highly complicated regularizer. In this paper, we propose a novel Multi-Layer Feature reduction method (MLFre) to quickly identify the inactive nodes (the groups of features with zero coefficients in the solution) hierarchically in a top-down fashion, which are guaranteed to be irrelevant to the response. Thus, we can remove the detected nodes from the optimization without sacrificing accuracy. The major challenge in developing such testing rules is due to the overlaps between the parents and their children nodes. By a novel hierarchical projection algorithm, MLFre is able to test the nodes independently from any of their ancestor nodes. Moreover, we can integrate MLFre---that has a low computational cost---with any existing solvers. Experiments on both synthetic and real data sets demonstrate that the speedup gained by MLFre can be orders of magnitude.

****

From random walks to distances on unweighted graphs

Tatsunori Hashimoto, Yi Sun, Tommi Jaakkola

Large unweighted directed graphs are commonly used to capture relations between entities. A fundamental problem in the analysis of such networks is to properly define the similarity or dissimilarity between any two vertices. Despite the significance of this problem, statistical characterization of the proposed metrics has been limited.We introduce and develop a class of techniques for analyzing random walks on graphs using stochastic calculus. Using these techniques we generalize results on the degeneracy of hitting times and analyze a metric based on the Laplace transformed hitting time (LTHT). The metric serves as a natural, provably well-behaved alternative to the expected hitting time. We establish a general correspondence between hitting times of the Brownian motion and analogous hitting times on the graph. We show that the LTHT is consistent with respect to the underlying metric of a geometric graph, preserves clustering tendency, and remains robust against random addition of non-geometric edges. Tests on simulated and real-world data show that the LTHT matches theoretical predictions and outperforms alternatives.

****

Tensorizing Neural Networks

Alexander Novikov, Dmitrii Podoprikhin, Anton Osokin, Dmitry P. Vetrov

Deep neural networks currently demonstrate state-of-the-art performance in several domains.At the same time, models of this class are very demanding in terms of computational resources. In particular, a large amount of memory is required by commonly used fully-connected layers, making it hard to use the models on low-end devices and stopping the further increase of the model size. In this paper we convert the dense weight matrices of the fully-connected layers to the Tensor Train format such that the number of parameters is reduced by a huge factor and at the same time the expressive power of the layer is preserved.In particular, for the Very Deep VGG networks we report the compression factor of the dense weight matrix of a fully-connected layer up to 200000 times leading to the compression factor of the whole network up to 7 times.

****

On some provably correct cases of variational inference for topic models

Pranjal Awasthi, Andrej Risteski

Variational inference is an efficient, popular heuristic used in the context of latent variable models. We provide the first analysis of instances where variational inference algorithms converge to the global optimum, in the setting of topic models. Our initializations are natural, one of them being used in LDA-c, the mostpopular implementation of variational inference.In addition to providing intuition into why this heuristic might work in practice, the multiplicative, rather than additive nature of the variational inference updates forces us to usenon-standard proof arguments, which we believe might be of general theoretical interest.

****

GAP Safe screening rules for sparse multi-task and multi-class models

Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, Joseph Salmon

High dimensional regression benefits from sparsity promoting regularizations. Screening rules leverage the known sparsity of the solution by ignoring some variables in the optimization, hence speeding up solvers. When the procedure is proven not to discard features wrongly the rules are said to be safe. In this paper we derive new safe rules for generalized linear models regularized with L1 and L1/L2 norms. The rules are based on duality gap computations and spherical safe regions whose diameters converge to zero. This allows to discard safely more variables, in particular for low regularization parameters. The GAP Safe rule can cope with any iterative solver and we illustrate its performance on coordinate descent for multi-task Lasso, binary and multinomial logistic regression, demonstrating significant speed ups on all tested datasets with respect to previous safe rules.

****

The Pareto Regret Frontier for Bandits

Tor Lattimore

Given a multi-armed bandit problem it may be desirable to achieve a smaller-than-usual worst-case regret for some special actions. I show that the price for such unbalanced worst-case regret guarantees is rather high. Specifically, if an algorithm enjoys a worst-case regret of B with respect to some action, then there must exist another action for which the worst-case regret is at least $\Omega(nK/B)$, where n is the horizon and K the number of actions. I also give upper bounds in both the stochastic and adversarial settings showing that this result cannot be improved. For the stochastic case the pareto regret frontier is characterised exactly up to constant factors.

****

Measuring Sample Quality with Stein's Method

Jackson Gorham, Lester Mackey

To improve the efficiency of Monte Carlo estimation, practitioners are turning to biased Markov chain Monte Carlo procedures that trade off asymptotic exactness for computational speed. The reasoning is sound: a reduction in variance due to more rapid sampling can outweigh the bias introduced. However, the inexactness creates new challenges for sampler and parameter selection, since standard measures of sample quality like effective sample size do not account for asymptotic bias. To address these challenges, we introduce a new computable quality measure based on Stein's method that bounds the discrepancy between sample and target expectations over a large class of test functions. We use our tool to compare exact, biased, and deterministic sample sequences and illustrate applications to hyperparameter selection, convergence rate assessment, and quantifying bias-variance tradeoffs in posterior inference.

****

Predtron: A Family of Online Algorithms for General Prediction Problems

Prateek Jain, Nagarajan Natarajan, Ambuj Tewari

Modern prediction problems arising in multilabel learning and learning to rank pose unique challenges to the classical theory of supervised learning. These problems have large prediction and label spaces of a combinatorial nature and involve sophisticated loss functions. We offer a general framework to derive mistake driven online algorithms and associated loss bounds. The key ingredients in our framework are a general loss function, a general vector space representation of

predictions, and a notion of margin with respect to a general norm. Our general algorithm, Predtron, yields the perceptron algorithm and its variants when instantiated on classic problems such as binary classification, multiclass classification, ordinal regression, and multilabel classification. For multilabel ranking and subset ranking, we derive novel algorithms, notions of margins, and loss bounds. A simulation study confirms the behavior predicted by our bounds and demonstrates the flexibility of the design choices in our framework.
****

MCMC for Variationally Sparse Gaussian Processes
James Hensman, Alexander G. Matthews, Maurizio Filippone, Zoubin Ghahramani
Gaussian process (GP) models form a core part of probabilistic machine learning. Considerable research effort has been made into attacking three issues with GP models: how to compute efficiently when the number of data is large; how to approximate the posterior when the likelihood is not Gaussian and how to estimate covariance function parameter posteriors. This paper simultaneously addresses these, using a variational approximation to the posterior which is sparse in sup- port of the function but otherwise free-form. The result is a Hybrid Monte-Carlo sampling scheme which allows for a non-Gaussian approximation over the function values and covariance parameters simultaneously, with efficient computations based on inducing-point sparse GPs.
****

Action-Conditional Video Prediction using Deep Networks in Atari Games
Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L. Lewis, Satinder Singh
Motivated by vision-based reinforcement learning (RL) problems, in particular Atari games from the recent benchmark Aracade Learning Environment (ALE), we consider spatio-temporal prediction problems where future (image-)frames are dependent on control variables or actions as well as previous frames. While not composed of natural scenes, frames in Atari games are high-dimensional in size, can involve tens of objects with one or more objects being controlled by the actions directly and many other objects being influenced indirectly, can involve entry and departure of objects, and can involve deep partial observability. We propose and evaluate two deep neural network architectures that consist of encoding, action-conditional transformation, and decoding layers based on convolutional neural networks and recurrent neural networks. Experimental results show that the proposed architectures are able to generate visually-realistic frames that are also useful for control over approximately 100-step action-conditional futures in some games. To the best of our knowledge, this paper is the first to make and evaluate long-term predictions on high-dimensional video conditioned by control inputs.
****

Unified View of Matrix Completion under General Structural Constraints
Suriya Gunasekar, Arindam Banerjee, Joydeep Ghosh
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

When are Kalman-Filter Restless Bandits Indexable?
Christopher R. Dance, Tomi Silander
We study the restless bandit associated with an extremely simple scalar Kalman filter model in discrete time. Under certain assumptions, we prove that the problem is {\it indexable} in the sense that the {\it Whittle index} is a non-decreasing function of the relevant belief state. In spite of the long history of this problem, this appears to be the first such proof. We use results about {\it Schur-convexity} and {\it mechanical words}, which are particularbinary strings intimately related to {\it palindromes}.
****

3D Object Proposals for Accurate Object Class Detection
Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G. Berneshawi, Huimin Ma, Sanja Fidler, Raquel Urtasun
The goal of this paper is to generate high-quality 3D object proposals in the co

ntext of autonomous driving. Our method exploits stereo imagery to place proposals in the form of 3D bounding boxes. We formulate the problem as minimizing an energy function encoding object size priors, ground plane as well as several depth informed features that reason about free space, point cloud densities and distance to the ground. Our experiments show significant performance gains over existing RGB and RGB-D object proposal methods on the challenging KITTI benchmark. Combined with convolutional neural net (CNN) scoring, our approach outperforms all existing results on all three KITTI object classes.

****

Interpolating Convex and Non-Convex Tensor Decompositions via the Subspace Norm
Qinqing Zheng, Ryota Tomioka

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Biologically Inspired Dynamic Textures for Probing Motion Perception
Jonathan Vacher, Andrew Isaac Meso, Laurent U. Perrinet, Gabriel Peyré

Perception is often described as a predictive process based on an optimal inference with respect to a generative model. We study here the principled construction of a generative model specifically crafted to probe motion perception. In that context, we first provide an axiomatic, biologically-driven derivation of the model. This model synthesizes random dynamic textures which are defined by stationary Gaussian distributions obtained by the random aggregation of warped patterns. Importantly, we show that this model can equivalently be described as a stochastic partial differential equation. Using this characterization of motion in images, it allows us to recast motion-energy models into a principled Bayesian inference framework. Finally, we apply these textures in order to psychophysically probe speed perception in humans. In this framework, while the likelihood is derived from the generative model, the prior is estimated from the observed results and accounts for the perceptual bias in a principled fashion.

****

Covariance-Controlled Adaptive Langevin Thermostat for Large-Scale Bayesian Sampling
Xiaocheng Shang, Zhanxing Zhu, Benedict Leimkuhler, Amos J. Storkey

Monte Carlo sampling for Bayesian posterior inference is a common approach used in machine learning. The Markov Chain Monte Carlo procedures that are used are often discrete-time analogues of associated stochastic differential equations (SDEs). These SDEs are guaranteed to leave invariant the required posterior distribution. An area of current research addresses the computational benefits of stochastic gradient methods in this setting. Existing techniques rely on estimating the variance or covariance of the subsampling error, and typically assume constant variance. In this article, we propose a covariance-controlled adaptive Langevin thermostat that can effectively dissipate parameter-dependent noise while maintaining a desired target distribution. The proposed method achieves a substantial speedup over popular alternative schemes for large-scale machine learning applications.

****

Semi-supervised Sequence Learning
Andrew M. Dai, Quoc V. Le

We present two approaches to use unlabeled data to improve Sequence Learningwith recurrent networks. The first approach is to predict what comes next in asequence, which is a language model in NLP. The second approach is to use asequence autoencoder, which reads the input sequence into a vector and predictsthe input sequence again. These two algorithms can be used as a “pretraining”algorithm for a later supervised sequence learning algorithm. In other words, theparameters obtained from the pretraining step can then be used as a starting pointfor other supervised training models. In our experiments, we find that long shortterm memory recurrent networks after pretrained with the two approaches becomemore stable to train and generalize better. With pretraining, we were able toachieve strong p erformance in many classification tasks, such as text classificationwith IMDB, DBpedia or image recognition in CIFAR-10.

****

Non-convex Statistical Optimization for Sparse Tensor Graphical Model

Wei Sun, Zhaoran Wang, Han Liu, Guang Cheng

We consider the estimation of sparse graphical models that characterize the dependency structure of high-dimensional tensor-valued data. To facilitate the estimation of the precision matrix corresponding to each way of the tensor, we assume the data follow a tensor normal distribution whose covariance has a Kronecker product structure. The penalized maximum likelihood estimation of this model involves minimizing a non-convex objective function. In spite of the non-convexity of this estimation problem, we prove that an alternating minimization algorithm, which iteratively estimates each sparse precision matrix while fixing the others, attains an estimator with the optimal statistical rate of convergence as well as consistent graph recovery. Notably, such an estimator achieves estimation consistency with only one tensor sample, which is unobserved in previous work. Our theoretical results are backed by thorough numerical studies.

****

Lifted Symmetry Detection and Breaking for MAP Inference

Timothy Kopp, Parag Singla, Henry Kautz

Symmetry breaking is a technique for speeding up propositional satisfiability testing by adding constraints to the theory that restrict the search space while preserving satisfiability. In this work, we extend symmetry breaking to the problem of model finding in weighted and unweighted relational theories, a class of problems that includes MAP inference in Markov Logic and similar statistical-relational languages. We introduce term symmetries, which are induced by an evidence set and extend to symmetries over a relational theory. We provide the important special case of term equivalent symmetries, showing that such symmetries can be found in low-degree polynomial time. We show how to break an exponential number of these symmetries with added constraints whose number is linear in the size of the domain. We demonstrate the effectiveness of these techniques through experiments in two relational domains. We also discuss the connections between relational symmetry breaking and work on lifted inference in statistical-relational reasoning.

****

Private Graphon Estimation for Sparse Graphs

Christian Borgs, Jennifer Chayes, Adam Smith

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Online Learning with Adversarial Delays

Kent Quanrud, Daniel Khashabi

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Solving Random Quadratic Systems of Equations Is Nearly as Easy as Solving Linear Systems

Yuxin Chen, Emmanuel Candes

This paper is concerned with finding a solution x to a quadratic system of equations yi = |< ai, x >|^2, i = 1, 2, ..., m. We prove that it is possible to solve unstructured quadratic systems in n variables exactly from O(n) equations in linear time, that is, in time proportional to reading and evaluating the data. This is accomplished by a novel procedure, which starting from an initial guess given by a spectral initialization procedure, attempts to minimize a non-convex objective. The proposed algorithm distinguishes from prior approaches by regularizing the initialization and descent procedures in an adaptive fashion, which disc ard terms bearing too much influence on the initial estimate or search directions. These careful selection rules---which effectively serve as a variance reduction scheme---provide a tighter initial guess, more robust descent directions, and thus enhanced practical performance. Further, this procedure also achieves a near-optimal statistical accuracy in the presence of noise. Finally, we demonstrate empirically that the computational cost of our algorithm is about four times that of solving a least-squares problem of the same size.

****

Statistical Topological Data Analysis - A Kernel Perspective

Roland Kwitt, Stefan Huber, Marc Niethammer, Weili Lin, Ulrich Bauer

We consider the problem of statistical computations with persistence diagrams, a summary representation of topological features in data. These diagrams encode persistent homology, a widely used invariant in topological data analysis. While several avenues towards a statistical treatment of the diagrams have been explored recently, we follow an alternative route that is motivated by the success of methods based on the embedding of probability measures into reproducing kernel Hilbert spaces. In fact, a positive definite kernel on persistence diagrams has recently been proposed, connecting persistent homology to popular kernel-based learning techniques such as support vector machines. However, important properties of that kernel which would enable a principled use in the context of probability measure embeddings remain to be explored. Our contribution is to close this gap by proving universality of a variant of the original kernel, and to demonstrate its effective use in two-sample hypothesis testing on synthetic as well as real-world data.

****

A Structural Smoothing Framework For Robust Graph Comparison

Pinar Yanardag, S.V.N. Vishwanathan

In this paper, we propose a general smoothing framework for graph kernels by taking \textit{structural similarity} into account, and apply it to derive smoothed variants of popular graph kernels. Our framework is inspired by state-of-the-art smoothing techniques used in natural language processing (NLP). However, unlike NLP applications which primarily deal with strings, we show how one can apply smoothing to a richer class of inter-dependent sub-structures that naturally arise in graphs. Moreover, we discuss extensions of the Pitman-Yor process that can be adapted to smooth structured objects thereby leading to novel graph kernels. Our kernels are able to tackle the diagonal dominance problem, while respecting the structural similarity between sub-structures,   especially under the presence of edge or label noise. Experimental evaluation shows that not only our kernels outperform the unsmoothed variants, but also achieve statistically significant improvements in classification accuracy over several other graph kernels that have been recently proposed in literature. Our kernels are competitive in terms of runtime, and offer a viable option for practitioners.

****

Bandits with Unobserved Confounders: A Causal Approach

Elias Bareinboim, Andrew Forney, Judea Pearl

The Multi-Armed Bandit problem constitutes an archetypal setting for sequential decision-making, permeating multiple domains including engineering, business, and medicine. One of the hallmarks of a bandit setting is the agent's capacity to explore its environment through active intervention, which contrasts with the ability to collect passive data by estimating associational relationships between actions and payouts. The existence of unobserved confounders, namely unmeasured variables affecting both the action and the outcome variables, implies that these two data-collection modes will in general not coincide. In this paper, we show that formalizing this distinction has conceptual and algorithmic implications to the bandit setting. The current generation of bandit algorithms implicitly try to maximize rewards based on estimation of the experimental distribution, which we show is not always the best strategy to pursue. Indeed, to achieve low regret in certain realistic classes of bandit problems (namely, in the face of unobserved confounders), both experimental and observational quantities are required by the rational agent. After this realization, we propose an optimization metric (employing both experimental and observational distributions) that bandit agents should pursue, and illustrate its benefits over traditional algorithms.

****

Scale Up Nonlinear Component Analysis with Doubly Stochastic Gradients

Bo Xie, Yingyu Liang, Le Song

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

A Market Framework for Eliciting Private Data

Bo Waggoner, Rafael Frongillo, Jacob D. Abernethy

We propose a mechanism for purchasing information from a sequence of participants.The participants may simply hold data points they wish to sell, or may have more sophisticated information; either way, they are incentivized to participate as long as they believe their data points are representative or their information will improve the mechanism's future prediction on a test set.The mechanism, which draws on the principles of prediction markets, has a bounded budget and minimizes generalization error for Bregman divergence loss functions.We then show how to modify this mechanism to preserve the privacy of participants' information: At any given time, the current prices and predictions of the mechanism reveal almost no information about any one participant, yet in total over all participants, information is accurately aggregated.

****

A Generalization of Submodular Cover via the Diminishing Return Property on the Integer Lattice

Tasuku Soma, Yuichi Yoshida

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Space-Time Local Embeddings

Ke Sun, Jun Wang, Alexandros Kalousis, Stephane Marchand-Maillet

Space-time is a profound concept in physics. This concept was shown to be useful for dimensionality reduction. We present basic definitions with interesting counter-intuitions. We give theoretical propositions to show that space-time is a more powerful representation than Euclidean space. We apply this concept to manifold learning for preserving local information. Empirical results on non-metric datasets show that more information can be preserved in space-time.

****

Mixing Time Estimation in Reversible Markov Chains from a Single Sample Path

Daniel J. Hsu, Aryeh Kontorovich, Csaba Szepesvari

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Online Rank Elicitation for Plackett-Luce: A Dueling Bandits Approach

Balázs Szörényi, Róbert Busa-Fekete, Adil Paul, Eyke Hüllermeier

We study the problem of online rank elicitation, assuming that rankings of a set of alternatives obey the Plackett-Luce distribution. Following the setting of the dueling bandits problem, the learner is allowed to query pairwise comparisons between alternatives, i.e., to sample pairwise marginals of the distribution in an online fashion. Using this information, the learner seeks to reliably predict the most probable ranking (or top-alternative). Our approach is based on constructing a surrogate probability distribution over rankings based on a sorting procedure, for which the pairwise marginals provably coincide with the marginals of the Plackett-Luce distribution. In addition to a formal performance and complexity analysis, we present first experimental studies.

****

Efficient Exact Gradient Update for training Deep Networks with Very Large Sparse Targets

Pascal Vincent, Alexandre de Brébisson, Xavier Bouthillier

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

A Gaussian Process Model of Quasar Spectral Energy Distributions

Andrew Miller, Albert Wu, Jeff Regier, Jon McAuliffe, Dustin Lang, Mr. Prabhat, David Schlegel, Ryan P. Adams

We propose a method for combining two sources of astronomical data, spectroscopy and photometry, that carry information about sources of light (e.g., stars, galaxies, and quasars) at extremely different spectral resolutions. Our model treats the spectral energy distribution (SED) of the radiation from a source as a latent variable that jointly explains both photometric and spectroscopic observations. We place a flexible, nonparametric prior over the SED of a light source that admits a physically interpretable decomposition, and allows us to tractably perform inference. We use our model to predict the distribution of the redshift of a quasar from five-band (low spectral resolution) photometric data, the so called ``photo-z'' problem. Our method shows that tools from machine learning and Bayesian statistics allow us to leverage multiple resolutions of information to make accurate predictions with well-characterized uncertainties.

****

Fast Convergence of Regularized Learning in Games

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, Robert E. Schapire

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Communication Complexity of Distributed Convex Learning and Optimization

Yossi Arjevani, Ohad Shamir

We study the fundamental limits to communication-efficient distributed methods for convex learning and optimization, under different assumptions on the information available to individual machines, and the types of functions considered. We identify cases where existing algorithms are already worst-case optimal, as well as cases where room for further improvement is still possible. Among other things, our results indicate that without similarity between the local objective functions (due to statistical data similarity or otherwise) many communication rounds may be required, even if the machines have unbounded computational power.

****

Large-Scale Bayesian Multi-Label Learning via Topic-Based Label Embeddings

Piyush Rai, Changwei Hu, Ricardo Henao, Lawrence Carin

We present a scalable Bayesian multi-label learning model based on learning low-dimensional label embeddings. Our model assumes that each label vector is generated as a weighted combination of a set of topics (each topic being a distribution over labels), where the combination weights (i.e., the embeddings) for each label vector are conditioned on the observed feature vector. This construction, coupled with a Bernoulli-Poisson link function for each label of the binary label vector, leads to a model with a computational cost that scales in the number of positive labels in the label matrix. This makes the model particularly appealing for real-world multi-label learning problems where the label matrix is usually very massive but highly sparse. Using a data-augmentation strategy leads to full local conjugacy in our model, facilitating simple and very efficient Gibbs sampling, as well as an Expectation Maximization algorithm for inference. Also, predicting the label vector at test time does not require doing an inference for the label embeddings and can be done in closed form. We report results on several benchmark data sets, comparing our model with various state-of-the art methods.

****

Probabilistic Line Searches for Stochastic Optimization

Maren Mahsereci, Philipp Hennig

In deterministic optimization, line searches are a standard tool ensuring stability and efficiency. Where only stochastic gradients are available, no direct equivalent has so far been formulated, because uncertain gradients do not allow for a strict sequence of decisions collapsing the search space. We construct a probabilistic line search by combining the structure of existing deterministic methods with notions from Bayesian optimization. Our method retains a Gaussian process surrogate of the univariate optimization objective, and uses a probabilistic belief over the Wolfe conditions to monitor the descent. The algorithm has very low computational cost, and no user-controlled parameters. Experiments show that it effectively removes the need to define a learning rate for stochastic gradient descent.

****

Sample Complexity of Learning Mahalanobis Distance Metrics

Nakul Verma, Kristin Branson

Metric learning seeks a transformation of the feature space that enhances prediction quality for a given task. In this work we provide PAC-style sample complexity rates for supervised metric learning. We give matching lower- and upper-bounds showing that sample complexity scales with the representation dimension when no assumptions are made about the underlying data distribution. In addition, by leveraging the structure of the data distribution, we provide rates fine-tuned to a specific notion of the intrinsic complexity of a given dataset, allowing us to relax the dependence on representation dimension. We show both theoretically and empirically that augmenting the metric learning optimization criterion with a simple norm-based regularization is important and can help adapt to a dataset’s intrinsic complexity yielding better generalization, thus partly explaining the empirical success of similar regularizations reported in previous works.

****

Sample Efficient Path Integral Control under Uncertainty

Yunpeng Pan, Evangelos Theodorou, Michail Kontitsis

We present a data-driven stochastic optimal control framework that is derived using the path integral (PI) control approach. We find iterative control laws analytically without a priori policy parameterization based on probabilistic representation of the learned dynamics model. The proposed algorithm operates in a forward-backward sweep manner which differentiate it from other PI-related methods that perform forward sampling to find open-loop optimal controls. Our method uses significantly less sampled data to find analytic control laws compared to other approaches within the PI control family that rely on extensive sampling from given dynamics models or trials on physical systems in a model-free fashion. In addition, the learned controllers can be generalized to new tasks without re-sampling based on the compositionality theory for the linearly-solvable optimal control framework.We provide experimental results on three different systems and comparisons with state-of-the-art model-based methods to demonstrate the efficiency and generalizability of the proposed framework.

****

Mind the Gap: A Generative Approach to Interpretable Feature Selection and Extraction

Been Kim, Julie A. Shah, Finale Doshi-Velez

We present the Mind the Gap Model (MGM), an approach for interpretable feature extraction and selection. By placing interpretability criteria directly into the model, we allow for the model to both optimize parameters related to interpretability and to directly report a global set of distinguishable dimensions to assist with further data exploration and hypothesis generation. MGM extracts distinguishing features on real-world datasets of animal features, recipes ingredients, and disease co-occurrence. It also maintains or improves performance when compared to related approaches. We perform a user study with domain experts to show the MGM's ability to help with dataset exploration.

****

Regularization Path of Cross-Validation Error Lower Bounds
Atsushi Shibagaki, Yoshiki Suzuki, Masayuki Karasuyama, Ichiro Takeuchi

Careful tuning of a regularization parameter is indispensable in many machine learning tasks because it has a significant impact on generalization performances. Nevertheless, current practice of regularization parameter tuning is more of an art than a science, e.g., it is hard to tell how many grid-points would be needed in cross-validation (CV) for obtaining a solution with sufficiently small CV error.In this paper we propose a novel framework for computing a lower bound of the CV errors as a function of the regularization parameter, which we call regularization path of CV error lower bounds.The proposed framework can be used for providing a theoretical approximation guarantee on a set of solutions in the sense that how far the CV error of the current best solution could be away from best possible CV error in the entire range of the regularization parameters.We demonstrate through numerical experiments that a theoretically guaranteed a choice of regularization parameter in the above sense is possible with reasonable computational costs.

****

Reflection, Refraction, and Hamiltonian Monte Carlo
Hadi Mohasel Afshar, Justin Domke

Hamiltonian Monte Carlo (HMC) is a successful approach for sampling from continuous densities. However, it has difficulty simulating Hamiltonian dynamics with non-smooth functions, leading to poor performance. This paper is motivated by the behavior of Hamiltonian dynamics in physical systems like optics. We introduce a modification of the Leapfrog discretization of Hamiltonian dynamics on piecewise continuous energies, where intersections of the trajectory with discontinuities are detected, and the momentum is reflected or refracted to compensate for the change in energy. We prove that this method preserves the correct stationary distribution when boundaries are affine. Experiments show that by reducing the number of rejected samples, this method improves on traditional HMC.

****

Exploring Models and Data for Image Question Answering
Mengye Ren, Ryan Kiros, Richard Zemel

This work aims to address the problem of image-based question-answering (QA) with new models and datasets. In our work, we propose to use neural networks and visual semantic embeddings, without intermediate stages such as object detection and image segmentation, to predict answers to simple questions about images. Our model performs 1.8 times better than the only published results on an existing image QA dataset. We also present a question generation algorithm that converts image descriptions, which are widely available, into QA form. We used this algorithm to produce an order-of-magnitude larger dataset, with more evenly distributed answers. A suite of baseline results on this new dataset are also presented.

****

Learning structured densities via infinite dimensional exponential families
Siqi Sun, Mladen Kolar, Jinbo Xu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Streaming Min-max Hypergraph Partitioning
Dan Alistarh, Jennifer Iglesias, Milan Vojnovic

In many applications, the data is of rich structure that can be represented by a hypergraph, where the data items are represented by vertices and the associations among items are represented by hyperedges. Equivalently, we are given an input bipartite graph with two types of vertices: items, and associations (which we refer to as topics). We consider the problem of partitioning the set of items into a given number of parts such that the maximum number of topics covered by a part of the partition is minimized. This is a natural clustering problem, with various applications, e.g. partitioning of a set of information objects such as documents, images, and videos, and load balancing in the context of computation p latforms.In this paper, we focus on the streaming computation model for this problem, in which items arrive online one at a time and each item must be assigned irrevocably to a part of the partition at its arrival time. Motivated by scalability requirements, we focus on the class of streaming computation algorithms with memory limited to be at most linear in the number of the parts of the partition. We show that a greedy assignment strategy is able to recover a hidden co-clustering of items under a natural set of recovery conditions. We also report results of an extensive empirical evaluation, which demonstrate that this greedy strategy yields superior performance when compared with alternative approaches.

****

Principal Differences Analysis: Interpretable Characterization of Differences between Distributions

Jonas W. Mueller, Tommi Jaakkola

We introduce principal differences analysis for analyzing differences between high-dimensional distributions. The method operates by finding the projection that maximizes the Wasserstein divergence between the resulting univariate populations. Relying on the Cramer-Wold device, it requires no assumptions about the form of the underlying distributions, nor the nature of their inter-class differences. A sparse variant of the method is introduced to identify features responsible for the differences. We provide algorithms for both the original minimax formulation as well as its semidefinite relaxation. In addition to deriving some convergence results, we illustrate how the approach may be applied to identify differences between cell populations in the somatosensory cortex and hippocampus as manifested by single cell RNA-seq. Our broader framework extends beyond the specific choice of Wasserstein divergence.

****

An Active Learning Framework using Sparse-Graph Codes for Sparse Polynomials and Graph Sketching

Xiao Li, Kannan Ramchandran

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Efficient Thompson Sampling for Online ■Matrix-Factorization Recommendation

Jaya Kawale, Hung H. Bui, Branislav Kveton, Long Tran-Thanh, Sanjay Chawla

Matrix factorization (MF) collaborative filtering is an effective and widely used method in recommendation systems. However, the problem of finding an optimal trade-off between exploration and exploitation (otherwise known as the bandit problem), a crucial problem in collaborative filtering from cold-start, has not been previously addressed.In this paper, we present a novel algorithm for online MF recommendation that automatically combines finding the most relevantitems with exploring new or less-recommended items.Our approach, called Particle Thompson Sampling for Matrix-Factorization, is based on the general Thompson sampling framework, but augmented with a novel efficient online Bayesian probabilistic matrix factorization method based on the Rao-Blackwellized particle filter.Extensive experiments in collaborative filtering using several real-world datasets demonstrate that our proposed algorithm significantly outperforms the current state-of-the-arts.

****

Structured Transforms for Small-Footprint Deep Learning

Vikas Sindhwani, Tara Sainath, Sanjiv Kumar

We consider the task of building compact deep learning pipelines suitable for deploymenton storage and power constrained mobile devices. We propose a uni-fied framework to learn a broad family of structured parameter matrices that arecharacterized by the notion of low displacement rank. Our structured transformsadmit fast function and gradient evaluation, and span a rich range of parametersharing configurations whose statistical modeling capacity can be explicitly tunedalong a continuum from structured to unstructured. Experimental results showthat these transforms can significantly accelerate inference and forward/backwardpasses du ring training, and offer superior accuracy-compactness-speed tradeoffsin comparison to a number of existing techniques. In keyword spotting applicationsin mobile speech recognition, our methods are much more effective thanstandard linear low-rank bottleneck layers and nearly retain the performance ofstate of the art models, while providing more than 3.5-fold compression.

****

Linear Multi-Resource Allocation with Semi-Bandit Feedback

Tor Lattimore, Koby Crammer, Csaba Szepesvari

We study an idealised sequential resource allocation problem. In each time step the learner chooses an allocation of several resource types between a number of tasks. Assigning more resources to a task increases the probability that it is completed. The problem is challenging because the alignment of the tasks to the resource types is unknown and the feedback is noisy. Our main contribution is the new setting and an algorithm with nearly-optimal regret analysis. Along the way we draw connections to the problem of minimising regret for stochastic linear bandits with heteroscedastic noise. We also present some new results for stochastic linear bandits on the hypercube that significantly out-performs existing work, especially in the sparse case.

****

On the Optimality of Classifier Chain for Multi-label Classification

Weiwei Liu, Ivor Tsang

To capture the interdependencies between labels in multi-label classification problems, classifier chain (CC) tries to take the multiple labels of each instance into account under a deterministic high-order Markov Chain model. Since its performance is sensitive to the choice of label order, the key issue is how to determine the optimal label order for CC. In this work, we first generalize the CC model over a random label order. Then, we present a theoretical analysis of the generalization error for the proposed generalized model. Based on our results, we propose a dynamic programming based classifier chain (CC-DP) algorithm to search the globally optimal label order for CC and a greedy classifier chain (CC-Greedy) algorithm to find a locally optimal CC. Comprehensive experiments on a number of real-world multi-label data sets from various domains demonstrate that our proposed CC-DP algorithm outperforms state-of-the-art approaches and the CC-Greedy algorithm achieves comparable prediction performance with CC-DP.

****

Consistent Multilabel Classification

Oluwasanmi O. Koyejo, Nagarajan Natarajan, Pradeep K. Ravikumar, Inderjit S. Dhillon

Multilabel classification is rapidly developing as an important aspect of modern predictive modeling, motivating study of its theoretical aspects. To this end, we propose a framework for constructing and analyzing multilabel classification metrics which reveals novel results on a parametric form for population optimal classifiers, and additional insight into the role of label correlations. In particular, we show that for multilabel metrics constructed as instance-, micro- and macro-averages, the population optimal classifier can be decomposed into binary classifiers based on the marginal instance-conditional distribution of each label, with a weak association between labels via the threshold. Thus, our analysis extends the state of the art from a few known multilabel classification metrics such as Hamming loss, to a general framework applicable to many of the classification metrics in common use. Based on the population-optimal classifier, we propose a computationally efficient and general-purpose plug-in classification algorithm, and prove its consistency with respect to the metric of interest. Empirical results on synthetic and benchmark datasets are supportive of our theoretical findings.

****

A Normative Theory of Adaptive Dimensionality Reduction in Neural Networks

Cengiz Pehlevan, Dmitri Chklovskii

To make sense of the world our brains must analyze high-dimensional datasets streamed by our sensory organs. Because such analysis begins with dimensionality reduction, modelling early sensory processing requires biologically plausible onli ne dimensionality reduction algorithms. Recently, we derived such an algorithm, termed similarity matching, from a Multidimensional Scaling (MDS) objective function. However, in the existing algorithm, the number of output dimensions is set a priori by the number of output neurons and cannot be changed. Because the number of informative dimensions in sensory inputs is variable there is a need for adaptive dimensionality reduction. Here, we derive biologically plausible dimensionality reduction algorithms which adapt the number of output dimensions to the eigenspectrum of the input covariance matrix. We formulate three objective functions which, in the offline setting, are optimized by the projections of the input dataset onto its principal subspace scaled by the eigenvalues of the output covariance matrix. In turn, the output eigenvalues are computed as i) soft-thresholded, ii) hard-thresholded, iii) equalized thresholded eigenvalues of the input covariance matrix. In the online setting, we derive the three corresponding adaptive algorithms and map them onto the dynamics of neuronal activity in networks with biologically plausible local learning rules. Remarkably, in the last two networks, neurons are divided into two classes which we identify with principal neurons and interneurons in biological circuits.

****

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison

Machine learning offers a fantastically powerful toolkit for building useful complexprediction systems quickly. This paper argues it is dangerous to think ofthese quick wins as coming for free. Using the software engineering frameworkof technical debt, we find it is common to incur massive ongoing maintenancecosts in real-world ML systems. We explore several ML-specific risk factors toaccount for in system design. These include boundary erosion, entanglement,hidden feedback loops, undeclared consumers, data dependencies, configurationissues, changes in the external world, and a variety of system-level anti-patterns.

****

NEXT: A System for Real-World Development, Evaluation, and Application of Active Learning

Kevin G. Jamieson, Lalit Jain, Chris Fernandez, Nicholas J. Glattard, Rob Nowak

Active learning methods automatically adapt data collection by selecting the most informative samples in order to accelerate machine learning. Because of this, real-world testing and comparing active learning algorithms requires collecting new datasets (adaptively), rather than simply applying algorithms to benchmark datasets, as is the norm in (passive) machine learning research. To facilitate the development, testing and deployment of active learning for real applications, we have built an open-source software system for large-scale active learning research and experimentation. The system, called NEXT, provides a unique platform for real-world, reproducible active learning research. This paper details the challenges of building the system and demonstrates its capabilities with several experiments. The results show how experimentation can help expose strengths and weaknesses of active learning algorithms, in sometimes unexpected and enlightening ways.

****

A Pseudo-Euclidean Iteration for Optimal Recovery in Noisy ICA

James R. Voss, Mikhail Belkin, Luis Rademacher

Independent Component Analysis (ICA) is a popular model for blind signal separation. The ICA model assumes that a number of independent source signals are linearly mixed to form the observed signals. We propose a new algorithm, PEGI (for pseudo-Euclidean Gradient Iteration), for provable model recovery for ICA with Gaussian noise. The main technical innovation of the algorithm is to use a fixed point iteration in a pseudo-Euclidean (indefinite “inner product”) space. The use of this indefinite “inner product” resolves technical issues common to several existing algorithms for noisy ICA. This leads to an algorithm which is conceptually simple, efficient and accurate in testing.Our second contribution is combining PEGI with the analysis of objectives for optimal recovery in the noisy ICA model. It has been observed that the direct approach of demixing with the inverse o f the mixing matrix is suboptimal for signal recovery in terms of the natural Signal to Interference plus Noise Ratio (SINR) criterion. There have been several partial solutions proposed in the ICA literature. It turns out that any solution to the mixing matrix reconstruction problem can be used to construct an SINR-optimal ICA demixing, despite the fact that SINR itself cannot be computed from data. That allows us to obtain a practical and provably SINR-optimal recovery method for ICA with arbitrary Gaussian noise.

****

Learning Structured Output Representation using Deep Conditional Generative Models

Kihyuk Sohn, Honglak Lee, Xinchen Yan

Supervised deep learning has been successfully applied for many recognition problems in machine learning and computer vision. Although it can approximate a complex many-to-one function very well when large number of training data is provided, the lack of probabilistic inference of the current supervised deep learning methods makes it difficult to model a complex structured output representations. In this work, we develop a scalable deep conditional generative model for structured output variables using Gaussian latent variables. The model is trained efficiently in the framework of stochastic gradient variational Bayes, and allows a fast prediction using stochastic feed-forward inference. In addition, we provide novel strategies to build a robust structured prediction algorithms, such as recurrent prediction network architecture, input noise-injection and multi-scale prediction training methods. In experiments, we demonstrate the effectiveness of our proposed algorithm in comparison to the deterministic deep neural network counterparts in generating diverse but realistic output representations using stochastic inference. Furthermore, the proposed schemes in training methods and architecture design were complimentary, which leads to achieve strong pixel-level object segmentation and semantic labeling performance on Caltech-UCSD Birds 200 and the subset of Labeled Faces in the Wild dataset.

****

Estimating Mixture Models via Mixtures of Polynomials

Sida Wang, Arun Tejasvi Chaganty, Percy S. Liang

Mixture modeling is a general technique for making any simple model more expressive through weighted combination. This generality and simplicity in part explains the success of the Expectation Maximization (EM) algorithm, in which updates are easy to derive for a wide class of mixture models. However, the likelihood of a mixture model is non-convex, so EM has no known global convergence guarantees. Recently, method of moments approaches offer global guarantees for some mixture models, but they do not extend easily to the range of mixture models that exist. In this work, we present Polymom, an unifying framework based on method of moments in which estimation procedures are easily derivable, just as in EM. Polymom is applicable when the moments of a single mixture component are polynomials of the parameters. Our key observation is that the moments of the mixture model are a mixture of these polynomials, which allows us to cast estimation as a Generalized Moment Problem. We solve its relaxations using semidefinite optimization, and then extract parameters using ideas from computer algebra. This framework allows us to draw insights and apply tools from convex optimization, computer algebra and the theory of moments to study problems in statistical estimation. Simulations show good empirical performance on several models.

****

Online Learning with Gaussian Payoffs and Side Observations

Yifan Wu, András György, Csaba Szepesvari

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Gradient-free Hamiltonian Monte Carlo with Efficient Kernel Exponential Families

Heiko Strathmann, Dino Sejdinovic, Samuel Livingstone, Zoltan Szabo, Arthur Gretton

We propose Kernel Hamiltonian Monte Carlo (KMC), a gradient-free adaptive MCMC algorithm based on Hamiltonian Monte Carlo (HMC). On target densities where classical HMC is not an option due to intractable gradients, KMC adaptively learns the target's gradient structure by fitting an exponential family model in a Reproducing Kernel Hilbert Space. Computational costs are reduced by two novel efficient approximations to this gradient. While being asymptotically exact, KMC mimics HMC in terms of sampling efficiency, and offers substantial mixing improvements over state-of-the-art gradient free samplers. We support our claims with experimental studies on both toy and real-world applications, including Approximate Bayesian Computation and exact-approximate MCMC.

****

Approximating Sparse PCA from Incomplete Data

ABHISEK KUNDU, Petros Drineas, Malik Magdon-Ismail

We study how well one can recover sparse principal componentsof a data matrix using a sketch formed from a few of its elements. We show that for a wide class of optimization problems,if the sketch is close (in the spectral norm) to the original datamatrix, then one can recover a near optimal solution to the optimizationproblem by using the sketch. In particular, we use this approach toobtain sparse principal components and show that for \math{m} data pointsin \math{n} dimensions,\math{O(\epsilon^{-2}\tilde k\max{m,n})} elements gives an\math{\epsilon}-additive approximation to the sparse PCA problem(\math{\tilde k} is the stable rank of the data matrix).We demonstrate our algorithms extensivelyon image, text, biological and financial data.The results show that not only are we able to recover the sparse PCAs from the incomplete data, but by using our sparse sketch, the running timedrops by a factor of five or more.

****

Regularization-Free Estimation in Trace Regression with Symmetric Positive Semidefinite Matrices

Martin Slawski, Ping Li, Matthias Hein

Trace regression models have received considerable attention in the context of matrix completion, quantum state tomography, and compressed sensing. Estimation of the underlying matrix from regularization-based approaches promoting low-rankedness, notably nuclear norm regularization, have enjoyed great popularity. In this paper, we argue that such regularization may no longer be necessary if the underlying matrix is symmetric positive semidefinite (spd) and the design satisfies certain conditions. In this situation, simple least squares estimation subject to an spd constraint may perform as well as regularization-based approaches with a proper choice of regularization parameter, which entails knowledge of the noise level and/or tuning. By contrast, constrained least squaresestimation comes without any tuning parameter and may hence be preferred due to its simplicity.

****

Learning visual biases from human imagination

Carl Vondrick, Hamed Pirsiavash, Aude Oliva, Antonio Torralba

Although the human visual system can recognize many concepts under challengingconditions, it still has some biases. In this paper, we investigate whether wecan extract these biases and transfer them into a machine recognition system.We introduce a novel method that, inspired by well-known tools in humanpsychophysics, estimates the biases that the human visual system might use forrecognition, but in computer vision feature spaces.   Our experiments aresurprising, and suggest that classifiers from the human visual system can betransferred into a machine with some success. Since these classifiers seem tocapture favorable biases in the human visual system, we further present an SVMformulation that constrains the orientation of the SVM hyperplane to agree withthe bias from human visual system. Our results suggest that transferring thishuman bias into machines may help object recognition systems generalize acrossdatasets and perform better when very little training data is available.

****

End-To-End Memory Networks

Sainbayar Sukhbaatar, arthur szlam, Jason Weston, Rob Fergus

We introduce a neural network with a recurrent attention model over a possibly l arge external memory. The architecture is a form of Memory Network (Weston et al., 2015) but unlike the model in that work, it is trained end-to-end, and hence requires significantly less supervision during training, making it more generally applicable in realistic settings. It can also be seen as an extension of RNNsearch to the case where multiple computational steps (hops) are performed per output symbol. The flexibility of the model allows us to apply it to tasks as diverse as (synthetic) question answering and to language modeling. For the former our approach is competitive with Memory Networks, but with less supervision. For the latter, on the Penn TreeBank and Text8 datasets our approach demonstrates comparable performance to RNNs and LSTMs. In both cases we show that the key concept of multiple computational hops yields improved results.

****

Fast Distributed k-Center Clustering with Outliers on Massive Data

Gustavo Malkomes, Matt J. Kusner, Wenlin Chen, Kilian Q. Weinberger, Benjamin Moseley

Clustering large data is a fundamental problem with a vast number of applications. Due to the increasing size of data, practitioners interested in clustering have turned to distributed computation methods. In this work, we consider the widely used k-center clustering problem and its variant used to handle noisy data, k-center with outliers. In the noise-free setting we demonstrate how a previously-proposed distributed method is actually an O(1)-approximation algorithm, which accurately explains its strong empirical performance. Additionally, in the noisy setting, we develop a novel distributed algorithm that is also an O(1)-approximation. These algorithms are highly parallel and lend themselves to virtually any distributed computing framework. We compare both empirically against the best known noisy sequential clustering methods and show that both distributed algorithms are consistently close to their sequential versions. The algorithms are all one can hope for in distributed settings: they are fast, memory efficient and they match their sequential counterparts.

****

BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions

Dominik Rothenhäusler, Christina Heinze, Jonas Peters, Nicolai Meinshausen

We propose a simple method to learn linear causal cyclic models in the presence of latent variables. The method relies on equilibrium data of the model recorded under a specific kind of interventions (``shift interventions''). The location and strength of these interventions do not have to be known and can be estimated from the data. Our method, called BACKSHIFT, only uses second moments of the data and performs simple joint matrix diagonalization, applied to differences between covariance matrices. We give a sufficient and necessary condition for identifiability of the system, which is fulfilled almost surely under some quite general assumptions if and only if there are at least three distinct experimental settings, one of which can be pure observational data. We demonstrate the performance on some simulated data and applications in flow cytometry and financial time series.

****

Lifelong Learning with Non-i.i.d. Tasks

Anastasia Pentina, Christoph H. Lampert

In this work we aim at extending theoretical foundations of lifelong learning. Previous work analyzing this scenario is based on the assumption that the tasks are sampled i.i.d. from a task environment or limited to strongly constrained data distributions. Instead we study two scenarios when lifelong learning is possible, even though the observed tasks do not form an i.i.d. sample: first, when they are sampled from the same environment, but possibly with dependencies, and second, when the task environment is allowed to change over time. In the first case we prove a PAC-Bayesian theorem, which can be seen as a direct generalization of the analogous previous result for the i.i.d. case. For the second scenario we propose to learn an inductive bias in form of a transfer procedure. We present a generalization bound and show on a toy example how it can be used to identify a beneficial transfer algorithm.

****

Regularized EM Algorithms: A Unified Framework and Statistical Guarantees
Xinyang Yi, Constantine Caramanis

Latent models are a fundamental modeling tool in machine learning applications, but they present significant computational and analytical challenges. The popular EM algorithm and its variants, is a much used algorithmic tool; yet our rigorous understanding of its performance is highly incomplete. Recently, work in [1] has demonstrated that for an important class of problems, EM exhibits linear local convergence. In the high-dimensional setting, however, the M-step may not be well defined. We address precisely this setting through a unified treatment using regularization. While regularization for high-dimensional problems is by now well understood, the iterative EM algorithm requires a careful balancing of making progress towards the solution while identifying the right structure (e.g., sparsity or low-rank). In particular, regularizing the M-step using the state-of-the-art high-dimensional prescriptions (e.g., `a la [19]) is not guaranteed to provide this balance. Our algorithm and analysis are linked in a way that reveals the balance between optimization and statistical errors. We specialize our general framework to sparse gaussian mixture models, high-dimensional mixed regression, and regression with missing variables, obtaining statistical guarantees for each of these examples.

****

Beyond Convexity: Stochastic Quasi-Convex Optimization
Elad Hazan, Kfir Levy, Shai Shalev-Shwartz

This poster has been moved from Monday #86 to Thursday #101.

****

Learning From Small Samples: An Analysis of Simple Decision Heuristics
Özgür ■im■ek, Marcus Buckmann

Simple decision heuristics are models of human and animal behavior that use few pieces of information---perhaps only a single piece of information---and integrate the pieces in simple ways, for example, by considering them sequentially, one at a time, or by giving them equal weight. It is unknown how quickly these heuristics can be learned from experience. We show, analytically and empirically, that only a few training samples lead to substantial progress in learning. We focus on three families of heuristics: single-cue decision making, lexicographic decision making, and tallying. Our empirical analysis is the most extensive to date, employing 63 natural data sets on diverse subjects.

****

Deep Temporal Sigmoid Belief Networks for Sequence Modeling
Zhe Gan, Chunyuan Li, Ricardo Henao, David E. Carlson, Lawrence Carin

Deep dynamic generative models are developed to learn sequential dependencies in time-series data. The multi-layered model is designed by constructing a hierarchy of temporal sigmoid belief networks (TSBNs), defined as a sequential stack of sigmoid belief networks (SBNs). Each SBN has a contextual hidden state, inherited from the previous SBNs in the sequence, and is used to regulate its hidden bias. Scalable learning and inference algorithms are derived by introducing a recognition model that yields fast sampling from the variational posterior. This recognition model is trained jointly with the generative model, by maximizing its variational lower bound on the log-likelihood. Experimental results on bouncing balls, polyphonic music, motion capture, and text streams show that the proposed approach achieves state-of-the-art predictive performance, and has the capacity to synthesize various sequences.

****

Subsampled Power Iteration: a Unified Algorithm for Block Models and Planted CSP's
Vitaly Feldman, Will Perkins, Santosh Vempala

We present an algorithm for recovering planted solutions in two well-known models, the stochastic block model and planted constraint satisfaction problems (CSP), via a common generalization in terms of random bipartite graphs. Our algorithm matches up to a constant factor the best-known bounds for the number of edges (or constraints) needed for perfect recovery and its running time is linear in the number of edges used. The time complexity is significantly better than both s pectral and SDP-based approaches.The main contribution of the algorithm is in the case of unequal sizes in the bipartition that arises in our reduction from the planted CSP. Here our algorithm succeeds at a significantly lower density than the spectral approaches, surpassing a barrier based on the spectral norm of a random matrix.Other significant features of the algorithm and analysis include (i) the critical use of power iteration with subsampling, which might be of independent interest; its analysis requires keeping track of multiple norms of an evolving solution (ii) the algorithm can be implemented statistically, i.e., with very limited access to the input distribution (iii) the algorithm is extremely simple to implement and runs in linear time, and thus is practical even for very large instances.

****

Learning Stationary Time Series using Gaussian Processes with Nonparametric Kernels

Felipe Tobar, Thang D. Bui, Richard E. Turner

We introduce the Gaussian Process Convolution Model (GPCM), a two-stage nonparametric generative procedure to model stationary signals as the convolution between a continuous-time white-noise process and a continuous-time linear filter drawn from Gaussian process. The GPCM is a continuous-time nonparametric-window moving average process and, conditionally, is itself a Gaussian process with a nonparametric kernel defined in a probabilistic fashion. The generative model can be equivalently considered in the frequency domain, where the power spectral density of the signal is specified using a Gaussian process. One of the main contributions of the paper is to develop a novel variational free-energy approach based on inter-domain inducing variables that efficiently learns the continuous-time linear filter and infers the driving white-noise process. In turn, this scheme provides closed-form probabilistic estimates of the covariance kernel and the noise-free signal both in denoising and prediction scenarios. Additionally, the variational inference procedure provides closed-form expressions for the approximate posterior of the spectral density given the observed data, leading to new Bayesian nonparametric approaches to spectrum estimation. The proposed GPCM is validated using synthetic and real-world signals.

****

Improved Iteration Complexity Bounds of Cyclic Block Coordinate Descent for Convex Problems

Ruoyu Sun, Mingyi Hong

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Community Detection via Measure Space Embedding

Mark Kozdoba, Shie Mannor

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Color Constancy by Learning to Predict Chromaticity from Luminance

Ayan Chakrabarti

Color constancy is the recovery of true surface color from observed color, and requires estimating the chromaticity of scene illumination to correct for the bias it induces. In this paper, we show that the per-pixel color statistics of natural scenes---without any spatial or semantic context---can by themselves be a powerful cue for color constancy. Specifically, we describe an illuminant estimation method that is built around a classifier for identifying the true chromaticity of a pixel given its luminance (absolute brightness across color channels). During inference, each pixel's observed color restricts its true chromaticity to those values that can be explained by one of a candidate set of illuminants, and applying the classifier over these values yields a distribution over the corresp onding illuminants. A global estimate for the scene illuminant is computed through a simple aggregation of these distributions across all pixels. We begin by simply defining the luminance-to-chromaticity classifier by computing empirical histograms over discretized chromaticity and luminance values from a training set of natural images. These histograms reflect a preference for hues corresponding to smooth reflectance functions, and for achromatic colors in brighter pixels. Despite its simplicity, the resulting estimation algorithm outperforms current state-of-the-art color constancy methods. Next, we propose a method to learn the luminance-to-chromaticity classifier end-to-end. Using stochastic gradient descent, we set chromaticity-luminance likelihoods to minimize errors in the final scene illuminant estimates on a training set. This leads to further improvements in accuracy, most significantly in the tail of the error distribution.

****

Sample Complexity Bounds for Iterative Stochastic Policy Optimization

Marin Kobilarov

This paper is concerned with robustness analysis of decision making under uncertainty. We consider a class of iterative stochastic policy optimization problems and analyze the resulting expected performance for each newly updated policy at each iteration. In particular, we employ concentration-of-measure inequalities to compute future expected cost and probability of constraint violation using empirical runs. A novel inequality bound is derived that accounts for the possibly unbounded change-of-measure likelihood ratio resulting from iterative policy adaptation. The bound serves as a high-confidence certificate for providing future performance or safety guarantees. The approach is illustrated with a simple robot control scenario and initial steps towards applications to challenging aerial vehicle navigation problems are presented.

****

Copeland Dueling Bandits

Masrour Zoghi, Zohar S. Karnin, Shimon Whiteson, Maarten de Rijke

A version of the dueling bandit problem is addressed in which a Condorcet winner may not exist. Two algorithms are proposed that instead seek to minimize regret with respect to the Copeland winner, which, unlike the Condorcet winner, is guaranteed to exist. The first, Copeland Confidence Bound (CCB), is designed for small numbers of arms, while the second, Scalable Copeland Bandits (SCB), works better for large-scale problems. We provide theoretical results bounding the regret accumulated by CCB and SCB, both substantially improving existing results. Such existing results either offer bounds of the form O(K log T) but require restrictive assumptions, or offer bounds of the form O(K^2 log T) without requiring such assumptions. Our results offer the best of both worlds: O(K log T) bounds without restrictive assumptions.

****

Taming the Wild: A Unified Analysis of Hogwild-Style Algorithms

Christopher M. De Sa, Ce Zhang, Kunle Olukotun, Christopher Ré, Christopher Ré

Stochastic gradient descent (SGD) is a ubiquitous algorithm for a variety of machine learning problems. Researchers and industry have developed several techniques to optimize SGD's runtime performance, including asynchronous execution and reduced precision. Our main result is a martingale-based analysis that enables us to capture the rich noise models that may arise from such techniques. Specifically, we useour new analysis in three ways: (1) we derive convergence rates for the convex case (Hogwild) with relaxed assumptions on the sparsity of the problem; (2) we analyze asynchronous SGD algorithms for non-convex matrix problems including matrix completion; and (3) we design and analyze an asynchronous SGD algorithm, called Buckwild, that uses lower-precision arithmetic. We show experimentally that our algorithms run efficiently for a variety of problems on modern hardware.

****

High-dimensional neural spike train analysis with generalized count linear dynamical systems

Yuanjun Gao, Lars Busing, Krishna V. Shenoy, John P. Cunningham

Latent factor models have been widely used to analyze simultaneous recordings of spike trains from large, heterogeneous neural populations. These models assume the signal of interest in the population is a low-dimensional latent intensity that evolves over time, which is observed in high dimension via noisy point-process observations. These techniques have been well used to capture neural correlations across a population and to provide a smooth, denoised, and concise representation of high-dimensional spiking data. One limitation of many current models is that the observation model is assumed to be Poisson, which lacks the flexibility to capture under- and over-dispersion that is common in recorded neural data, thereby introducing bias into estimates of covariance. Here we develop the generalized count linear dynamical system, which relaxes the Poisson assumption by using a more general exponential family for count data. In addition to containing Poisson, Bernoulli, negative binomial, and other common count distributions as special cases, we show that this model can be tractably learned by extending recent advances in variational inference techniques. We apply our model to data from primate motor cortex and demonstrate performance improvements over state-of-the-art methods, both in capturing the variance structure of the data and in held-out prediction.

****

Neural Adaptive Sequential Monte Carlo

Shixiang (Shane) Gu, Zoubin Ghahramani, Richard E. Turner

Sequential Monte Carlo (SMC), or particle filtering, is a popular class of methods for sampling from an intractable target distribution using a sequence of simpler intermediate distributions. Like other importance sampling-based methods, performance is critically dependent on the proposal distribution: a bad proposal can lead to arbitrarily inaccurate estimates of the target distribution. This paper presents a new method for automatically adapting the proposal using an approximation of the Kullback-Leibler divergence between the true posterior and the proposal distribution. The method is very flexible, applicable to any parameterized proposal distribution and it supports online and batch variants. We use the new framework to adapt powerful proposal distributions with rich parameterizations based upon neural networks leading to Neural Adaptive Sequential Monte Carlo (NASMC). Experiments indicate that NASMC significantly improves inference in a non-linear state space model outperforming adaptive proposal methods including the Extended Kalman and Unscented Particle Filters. Experiments also indicate that improved inference translates into improved parameter learning when NASMC is used as a subroutine of Particle Marginal Metropolis Hastings. Finally we show that NASMC is able to train a latent variable recurrent neural network (LV-RNN) achieving results that compete with the state-of-the-art for polymorphic music modelling. NASMC can be seen as bridging the gap between adaptive SMC methods and the recent work in scalable, black-box variational inference.

****

Supervised Learning for Dynamical System Learning

Ahmed Hefny, Carlton Downey, Geoffrey J. Gordon

Recently there has been substantial interest in spectral methods for learning dynamical systems. These methods are popular since they often offer a good tradeoffbetween computational and statistical efficiency. Unfortunately, they can be difficult to use and extend in practice: e.g., they can make it difficult to incorporateprior information such as sparsity or structure. To address this problem, we presenta new view of dynamical system learning: we show how to learn dynamical systems by solving a sequence of ordinary supervised learning problems, therebyallowing users to incorporate prior knowledge via standard techniques such asL1 regularization. Many existing spectral methods are special cases of this newframework, using linear regression as the supervised learner. We demonstrate theeffectiveness of our framework by showing examples where nonlinear regressionor lasso let us learn better state representations than plain linear regression does; the correctness of these instances follows directly from our general analysis.

****

A Complete Recipe for Stochastic Gradient MCMC

Yi-An Ma, Tianqi Chen, Emily Fox

Many recent Markov chain Monte Carlo (MCMC) samplers leverage continuous dynamic s to define a transition kernel that efficiently explores a target distribution. In tandem, a focus has been on devising scalable variants that subsample the data and use stochastic gradients in place of full-data gradients in the dynamic simulations. However, such stochastic gradient MCMC samplers have lagged behind their full-data counterparts in terms of the complexity of dynamics considered since proving convergence in the presence of the stochastic gradient noise is non-trivial. Even with simple dynamics, significant physical intuition is often required to modify the dynamical system to account for the stochastic gradient noise. In this paper, we provide a general recipe for constructing MCMC samplers--including stochastic gradient versions--based on continuous Markov processes specified via two matrices. We constructively prove that the framework is complete. That is, any continuous Markov process that provides samples from the target distribution can be written in our framework. We show how previous continuous-dynamic samplers can be trivially reinvented in our framework, avoiding the complicated sampler-specific proofs. We likewise use our recipe to straightforwardly propose a new state-adaptive sampler: stochastic gradient Riemann Hamiltonian Monte Carlo (SGRHMC). Our experiments on simulated data and a streaming Wikipedia analysis demonstrate that the proposed SGRHMC sampler inherits the benefits of Riemann HMC, with the scalability of stochastic gradient methods.

****

Segregated Graphs and Marginals of Chain Graph Models

Ilya Shpitser

Bayesian networks are a popular representation of asymmetric (for example causal) relationships between random variables. Markov random fields (MRFs) are a complementary model of symmetric relationships used in computer vision, spatial modeling, and social and gene expression networks. A chain graph model under the Lauritzen-Wermuth-Frydenberg interpretation (hereafter a chain graph model) generalizes both Bayesian networks and MRFs, and can represent asymmetric and symmetric relationships together.As in other graphical models, the set of marginals from distributions in a chain graph model induced by the presence of hidden variables forms a complex model. One recent approach to the study of marginal graphicals models is to consider a well-behaved supermodel. Such a supermodel of marginals of Bayesian networks, defined only by conditional independences, and termed the ordinary Markov model, was studied at length in (Evans and Richardson, 2014).In this paper, we show that special mixed graphs which we call segregated graphs can be associated, via a Markov property, with supermodels of a marginal of chain graphs defined only by conditional independences. Special features of segregated graphs imply the existence of a very natural factorization for these supermodels, and imply many existing results on the chain graph model, and ordinary Markov model carry over. Our results suggest that segregated graphs define an analogue of the ordinary Markov model for marginals of chain graph models.

****

Rethinking LDA: Moment Matching for Discrete ICA

Anastasia Podosinnikova, Francis Bach, Simon Lacoste-Julien

We consider moment matching techniques for estimation in Latent Dirichlet Allocation (LDA). By drawing explicit links between LDA and discrete versions of independent component analysis (ICA), we first derive a new set of cumulant-based tensors, with an improved sample complexity. Moreover, we reuse standard ICA techniques such as joint diagonalization of tensors to improve over existing methods based on the tensor power method. In an extensive set of experiments on both synthetic and real datasets, we show that our new combination of tensors and orthogonal joint diagonalization techniques outperforms existing moment matching methods.

****

Max-Margin Deep Generative Models

Chongxuan Li, Jun Zhu, Tianlin Shi, Bo Zhang

Deep generative models (DGMs) are effective on learning multilayered representations of complex data and performing inference of input data by exploring the generative ability. However, little work has been done on examining or empowering the discriminative ability of DGMs on making accurate predictions. This paper pre sents max-margin deep generative models (mmDGMs), which explore the strongly discriminative principle of max-margin learning to improve the discriminative power of DGMs, while retaining the generative capability. We develop an efficient doubly stochastic subgradient algorithm for the piecewise linear objective. Empirical results on MNIST and SVHN datasets demonstrate that (1) max-margin learning can significantly improve the prediction performance of DGMs and meanwhile retain the generative ability; and (2) mmDGMs are competitive to the state-of-the-art fully discriminative networks by employing deep convolutional neural networks (CNNs) as both recognition and generative models.

****

Convolutional Neural Networks with Intra-Layer Recurrent Connections for Scene Labeling

Ming Liang, Xiaolin Hu, Bo Zhang

Scene labeling is a challenging computer vision task. It requires the use of both local discriminative features and global context information. We adopt a deep recurrent convolutional neural network (RCNN) for this task, which is originally proposed for object recognition. Different from traditional convolutional neural networks (CNN), this model has intra-layer recurrent connections in the convolutional layers. Therefore each convolutional layer becomes a two-dimensional recurrent neural network. The units receive constant feed-forward inputs from the previous layer and recurrent inputs from their neighborhoods. While recurrent iterations proceed, the region of context captured by each unit expands. In this way, feature extraction and context modulation are seamlessly integrated, which is different from typical methods that entail separate modules for the two steps. To further utilize the context, a multi-scale RCNN is proposed. Over two benchmark datasets, Standford Background and Sift Flow, the model outperforms many state-of-the-art models in accuracy and efficiency.

****

Individual Planning in Infinite-Horizon Multiagent Settings: Inference, Structure and Scalability

Xia Qu, Prashant Doshi

This paper provides the first formalization of self-interested planning in multiagent settings using expectation-maximization (EM). Our formalization in the context of infinite-horizon and finitely-nested interactive POMDPs (I-POMDP) is distinct from EM formulations for POMDPs and cooperative multiagent planning frameworks. We exploit the graphical model structure specific to I-POMDPs, and present a new approach based on block-coordinate descent for further speed up. Forward filtering-backward sampling -- a combination of exact filtering with sampling -- is explored to exploit problem structure.

****

Expectation Particle Belief Propagation

Thibaut Lienart, Yee Whye Teh, Arnaud Doucet

We propose an original particle-based implementation of the Loopy Belief Propagation (LPB) algorithm for pairwise Markov Random Fields (MRF) on a continuous state space. The algorithm constructs adaptively efficient proposal distributions approximating the local beliefs at each note of the MRF. This is achieved by considering proposal distributions in the exponential family whose parameters are updated iterately in an Expectation Propagation (EP) framework. The proposed particle scheme provides consistent estimation of the LBP marginals as the number of particles increases. We demonstrate that it provides more accurate results than the Particle Belief Propagation (PBP) algorithm of Ihler and McAllester (2009) at a fraction of the computational cost and is additionally more robust empirically. The computational complexity of our algorithm at each iteration is quadratic in the number of particles. We also propose an accelerated implementation with sub-quadratic computational complexity which still provides consistent estimates of the loopy BP marginal distributions and performs almost as well as the original procedure.

****

Secure Multi-party Differential Privacy

Peter Kairouz, Sewoong Oh, Pramod Viswanath

We study the problem of multi-party interactive function computation under differential privacy. In this setting, each party is interested in computing a function on its private bit and all the other parties' bits. The function to be computed can vary from one party to the other. Moreover, there could be a central observer who is interested in computing a separate function on all the parties' bits. Differential privacy ensures that there remains an uncertainty in any party's bit even when given the transcript of interactions and all other parties' bits. Performance at each party is measured via the accuracy of the function to be computed. We allow for an arbitrary cost metric to measure the distortion between the true and the computed function values. Our main result is the optimality of a simple non-interactive protocol: each party randomizes its bit (sufficiently) and shares the privatized version with the other parties. This optimality result is very general: it holds for all types of functions, heterogeneous privacy conditions on the parties, all types of cost metrics, and both average and worst-case (over the inputs) measures of accuracy.

****

Embed to Control: A Locally Linear Latent Dynamics Model for Control from Raw Images

Manuel Watter, Jost Springenberg, Joschka Boedecker, Martin Riedmiller

We introduce Embed to Control (E2C), a method for model learning and control of non-linear dynamical systems from raw pixel images. E2C consists of a deep generative model, belonging to the family of variational autoencoders, that learns to generate image trajectories from a latent space in which the dynamics is constrained to be locally linear. Our model is derived directly from an optimal control formulation in latent space, supports long-term prediction of image sequences and exhibits strong performance on a variety of complex control problems.

****

Newton-Stein Method: A Second Order Method for GLMs via Stein's Lemma

Murat A. Erdogdu

We consider the problem of efficiently computing the maximum likelihood estimator in Generalized Linear Models (GLMs)when the number of observations is much larger than the number of coefficients (n > > p > > 1). In this regime, optimization algorithms can immensely benefit fromapproximate second order information.We propose an alternative way of constructing the curvature information by formulatingit as an estimation problem and applying a Stein-type lemma, which allows further improvements through sub-sampling andeigenvalue thresholding.Our algorithm enjoys fast convergence rates, resembling that of second order methods, with modest per-iteration cost. We provide its convergence analysis for the case where the rows of the design matrix are i.i.d. samples with bounded support.We show that the convergence has two phases, aquadratic phase followed by a linear phase. Finally,we empirically demonstrate that our algorithm achieves the highest performancecompared to various algorithms on several datasets.

****

Is Approval Voting Optimal Given Approval Votes?

Ariel D. Procaccia, Nisarg Shah

Some crowdsourcing platforms ask workers to express their opinions by approving a set of k good alternatives. It seems that the only reasonable way to aggregate these k-approval votes is the approval voting rule, which simply counts the number of times each alternative was approved. We challenge this assertion by proposing a probabilistic framework of noisy voting, and asking whether approval voting yields an alternative that is most likely to be the best alternative, given k-approval votes. While the answer is generally positive, our theoretical and empirical results call attention to situations where approval voting is suboptimal.

****

Optimization Monte Carlo: Efficient and Embarrassingly Parallel Likelihood-Free Inference

Ted Meeds, Max Welling

We describe an embarrassingly parallel, anytime Monte Carlo method for likelihood-free models. The algorithm starts with the view that the stochasticity of the pseudo-samples generated by the simulator can be controlled externally by a vec tor of random numbers u, in such a way that the outcome, knowing u, is deterministic. For each instantiation of u we run an optimization procedure to minimize the distance between summary statistics of the simulator and the data. After reweighing these samples using the prior and the Jacobian (accounting for the change of volume in transforming from the space of summary statistics to the space of parameters) we show that this weighted ensemble represents a Monte Carlo estimate of the posterior distribution. The procedure can be run embarrassingly parallel (each node handling one sample) and anytime (by allocating resources to the worst performing sample). The procedure is validated on six experiments.

****

Basis refinement strategies for linear value function approximation in MDPs

Gheorghe Comanici, Doina Precup, Prakash Panangaden

We provide a theoretical framework for analyzing basis function construction for linear value function approximation in Markov Decision Processes (MDPs). We show that important existing methods, such as Krylov bases and Bellman-error-based methods are a special case of the general framework we develop. We provide a general algorithmic framework for computing basis function refinements which “respect” the dynamics of the environment, and we derive approximation error bounds that apply for any algorithm respecting this general framework. We also show how, using ideas related to bisimulation metrics, one can translate basis refinement into a process of finding “prototypes” that are diverse enough to represent the given MDP.

****

StopWasting My Gradients: Practical SVRG

Reza Babanezhad Harikandeh, Mohamed Osama Ahmed, Alim Virani, Mark Schmidt, Jakub Kone■ný, Scott Sallinen

We present and analyze several strategies for improving the performance ofstochastic variance-reduced gradient (SVRG) methods. We first show that theconvergence rate of these methods can be preserved under a decreasing sequenceof errors in the control variate, and use this to derive variants of SVRG that usegrowing-batch strategies to reduce the number of gradient calculations requiredin the early iterations. We further (i) show how to exploit support vectors to reducethe number of gradient computations in the later iterations, (ii) prove that thecommonly–used regularized SVRG iteration is justified and improves the convergencerate, (iii) consider alternate mini-batch selection strategies, and (iv) considerthe generalization error of the method.

****

Saliency, Scale and Information: Towards a Unifying Theory

Shafin Rahman, Neil Bruce

In this paper we present a definition for visual saliency grounded in information theory. This proposal is shown to relate to a variety of classic research contributions in scale-space theory, interest point detection, bilateral filtering, and to existing models of visual saliency. Based on the proposed definition of visual saliency, we demonstrate results competitive with the state-of-the art for both prediction of human fixations, and segmentation of salient objects. We also characterize different properties of this model including robustness to image transformations, and extension to a wide range of other data types with 3D mesh models serving as an example. Finally, we relate this proposal more generally to the role of saliency computation in visual information processing and draw connections to putative mechanisms for saliency computation in human vision.

****

Efficient Learning of Continuous-Time Hidden Markov Models for Disease Progression

Yu-Ying Liu, Shuang Li, Fuxin Li, Le Song, James M. Rehg

The Continuous-Time Hidden Markov Model (CT-HMM) is an attractive approach to modeling disease progression due to its ability to describe noisy observations arriving irregularly in time. However, the lack of an efficient parameter learning algorithm for CT-HMM restricts its use to very small models or requires unrealistic constraints on the state transitions. In this paper, we present the first complete characterization of efficient EM-based learning methods for CT-HMM models

. We demonstrate that the learning problem consists of two challenges: the estimation of posterior state probabilities and the computation of end-state conditioned statistics. We solve the first challenge by reformulating the estimation problem in terms of an equivalent discrete time-inhomogeneous hidden Markov model. The second challenge is addressed by adapting three approaches from the continuous time Markov chain literature to the CT-HMM domain. We demonstrate the use of CT-HMMs with more than 100 states to visualize and predict disease progression using a glaucoma dataset and an Alzheimer's disease dataset.
****

Efficient Output Kernel Learning for Multiple Tasks
Pratik Kumar Jawanpuria, Maksim Lapin, Matthias Hein, Bernt Schiele
The paradigm of multi-task learning is that one can achieve better generalization by learning tasks jointly and thus exploiting the similarity between the tasks rather than learning them independently of each other. While previously the relationship between tasks had to be user-defined in the form of an output kernel, recent approaches jointly learn the tasks and the output kernel. As the output kernel is a positive semidefinite matrix, the resulting optimization problems are not scalable in the number of tasks as an eigendecomposition is required in each step. Using the theory of positive semidefinite kernels we show in this paper that for a certain class of regularizers on the output kernel, the constraint of being positive semidefinite can be dropped as it is automatically satisfied for the relaxed problem. This leads to an unconstrained dual problem which can be solved efficiently. Experiments on several multi-task and multi-class data sets illustrate the efficacy of our approach in terms of computational efficiency as well as generalization performance.
****

Texture Synthesis Using Convolutional Neural Networks
Leon Gatys, Alexander S. Ecker, Matthias Bethge
Here we introduce a new model of natural textures based on the feature spaces of convolutional neural networks optimised for object recognition. Samples from the model are of high perceptual quality demonstrating the generative power of neural networks trained in a purely discriminative fashion. Within the model, textures are represented by the correlations between feature maps in several layers of the network. We show that across layers the texture representations increasingly capture the statistical properties of natural images while making object information more and more explicit. The model provides a new tool to generate stimuli for neuroscience and might offer insights into the deep representations learned by convolutional neural networks.
****

Hessian-free Optimization for Learning Deep Multidimensional Recurrent Neural Networks
Minhyung Cho, Chandra Dhir, Jaehyung Lee
Multidimensional recurrent neural networks (MDRNNs) have shown a remarkable performance in the area of speech and handwriting recognition. The performance of an MDRNN is improved by further increasing its depth, and the difficulty of learning the deeper network is overcome by using Hessian-free (HF) optimization. Given that connectionist temporal classification (CTC) is utilized as an objective of learning an MDRNN for sequence labeling, the non-convexity of CTC poses a problem when applying HF to the network. As a solution, a convex approximation of CTC is formulated and its relationship with the EM algorithm and the Fisher information matrix is discussed. An MDRNN up to a depth of 15 layers is successfully trained using HF, resulting in an improved performance for sequence labeling.
****

Matrix Completion from Fewer Entries: Spectral Detectability and Rank Estimation
Alaa Saade, Florent Krzakala, Lenka Zdeborová
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Large-scale probabilistic predictors with and without guarantees of validity
Vladimir Vovk, Ivan Petej, Valentina Fedorova
This paper studies theoretically and empirically a method of turning machine-learning algorithms into probabilistic predictors that automatically enjoys a property of validity (perfect calibration) and is computationally efficient. The price to pay for perfect calibration is that these probabilistic predictors produce imprecise (in practice, almost precise for large data sets) probabilities. When these imprecise probabilities are merged into precise probabilities, the resulting predictors, while losing the theoretical property of perfect calibration, are consistently more accurate than the existing methods in empirical studies.
****
Learning with a Wasserstein Loss
Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, Tomaso A. Poggio
Learning to predict multi-label outputs is challenging, but in many problems there is a natural metric on the outputs that can be used to improve predictions. In this paper we develop a loss function for multi-label learning, based on the Wasserstein distance. The Wasserstein distance provides a natural notion of dissimilarity for probability measures. Although optimizing with respect to the exact Wasserstein distance is costly, recent work has described a regularized approximation that is efficiently computed. We describe an efficient learning algorithm based on this regularization, as well as a novel extension of the Wasserstein distance from probability measures to unnormalized measures. We also describe a statistical learning bound for the loss. The Wasserstein loss can encourage smoothness of the predictions with respect to a chosen metric on the output space. We demonstrate this property on a real-data tag prediction problem, using the Yahoo Flickr Creative Commons dataset, outperforming a baseline that doesn't use the metric.
****
Deep Generative Image Models using a ■Laplacian Pyramid of Adversarial Networks
Emily L. Denton, Soumith Chintala, arthur szlam, Rob Fergus
In this paper we introduce a generative model capable of producing high quality samples of natural images. Our approach uses a cascade of convolutional networks (convnets) within a Laplacian pyramid framework to generate images in a coarse-to-fine fashion. At each level of the pyramid a separate generative convnet model is trained using the Generative Adversarial Nets (GAN) approach. Samples drawn from our model are of significantly higher quality than existing models. In a quantitive assessment by human evaluators our CIFAR10 samples were mistaken for real images around 40% of the time, compared to 10% for GAN samples. We also show samples from more diverse datasets such as STL10 and LSUN.
****
Estimating Jaccard Index with Missing Observations: A Matrix Calibration Approach
Wenye Li
The Jaccard index is a standard statistics for comparing the pairwise similarity between data samples. This paper investigates the problem of estimating a Jaccard index matrix when there are missing observations in data samples. Starting from a Jaccard index matrix approximated from the incomplete data, our method calibrates the matrix to meet the requirement of positive semi-definiteness and other constraints, through a simple alternating projection algorithm. Compared with conventional approaches that estimate the similarity matrix based on the imputed data, our method has a strong advantage in that the calibrated matrix is guaranteed to be closer to the unknown ground truth in the Frobenius norm than the un-calibrated matrix (except in special cases they are identical). We carried out a series of empirical experiments and the results confirmed our theoretical justification. The evaluation also reported significantly improved results in real learning tasks on benchmarked datasets.
****
On Top-k Selection in Multi-Armed Bandits and Hidden Bipartite Graphs
Wei Cao, Jian Li, Yufei Tao, Zhize Li
Requests for name changes in the electronic proceedings will be accepted with no

questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Black-box optimization of noisy functions with unknown smoothness
Jean-Bastien Grill, Michal Valko, Remi Munos, Remi Munos

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding
Rie Johnson, Tong Zhang

This paper presents a new semi-supervised framework with convolutional neural networks (CNNs) for text categorization. Unlike the previous approaches that rely on word embeddings, our method learns embeddings of small text regions from unlabeled data for integration into a supervised CNN. The proposed scheme for embedding learning is based on the idea of two-view semi-supervised learning, which is intended to be useful for the task of interest even though the training is done on unlabeled data. Our models achieve better results than previous approaches on sentiment classification and topic classification tasks.

****

Fast Rates for Exp-concave Empirical Risk Minimization
Tomer Koren, Kfir Levy

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Learning both Weights and Connections for Efficient Neural Network
Song Han, Jeff Pool, John Tran, William Dally

Neural networks are both computationally intensive and memory intensive, making them difficult to deploy on embedded systems. Also, conventional networks fix the architecture before training starts; as a result, training cannot improve the architecture. To address these limitations, we describe a method to reduce the storage and computation required by neural networks by an order of magnitude without affecting their accuracy by learning only the important connections. Our method prunes redundant connections using a three-step method. First, we train the network to learn which connections are important. Next, we prune the unimportant connections. Finally, we retrain the network to fine tune the weights of the remaining connections. On the ImageNet dataset, our method reduced the number of parameters of AlexNet by a factor of 9×, from 61 million to 6.7 million, without incurring accuracy loss. Similar experiments with VGG-16 found that the total number of parameters can be reduced by 13×, from 138 million to 10.3 million, again with no loss of accuracy.

****

Bayesian dark knowledge
Anoop Korattikara Balan, Vivek Rathod, Kevin P. Murphy, Max Welling

We consider the problem of Bayesian parameter estimation for deep neural networks, which is important in problem settings where we may have little data, and/ or where we need accurate posterior predictive densities p(y|x, D), e.g., for applications involving bandits or active learning. One simple approach to this is to use online Monte Carlo methods, such as SGLD (stochastic gradient Langevin dynamics). Unfortunately, such a method needs to store many copies of the parameters (which wastes memory), and needs to make predictions using many versions of the model (which wastes time).We describe a method for “distilling” a Monte Carlo approximation to the posterior predictive density into a more compact form, namely a single deep neural network. We compare to two very recent approaches to Bayesian neural networks, namely an approach based on expectation propagation [HLA15

] and an approach based on variational Bayes [BCKW15]. Our method performs better than both of these, is much simpler to implement, and uses less computation at test time.

****

On the Convergence of Stochastic Gradient MCMC Algorithms with High-Order Integrators

Changyou Chen, Nan Ding, Lawrence Carin

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Teaching Machines to Read and Comprehend

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, Phil Blunsom

Teaching machines to read natural language documents remains an elusive challenge. Machine reading systems can be tested on their ability to answer questions posed on the contents of documents that they have seen, but until now large scale training and test datasets have been missing for this type of evaluation. In this work we define a new methodology that resolves this bottleneck and provides large scale supervised reading comprehension data. This allows us to develop a class of attention based deep neural networks that learn to read real documents and answer complex questions with minimal prior knowledge of language structure.

****

Synaptic Sampling: A Bayesian Approach to Neural Network Plasticity and Rewiring

David Kappel, Stefan Habenschuss, Robert Legenstein, Wolfgang Maass

We reexamine in this article the conceptual and mathematical framework for understanding the organization of plasticity in spiking neural networks. We propose that inherent stochasticity enables synaptic plasticity to carry out probabilistic inference by sampling from a posterior distribution of synaptic parameters. This view provides a viable alternative to existing models that propose convergence of synaptic weights to maximum likelihood parameters. It explains how priors on weight distributions and connection probabilities can be merged optimally with learned experience. In simulations we show that our model for synaptic plasticity allows spiking neural networks to compensate continuously for unforeseen disturbances. Furthermore it provides a normative mathematical framework to better understand the permanent variability and rewiring observed in brain networks.

****

Alternating Minimization for Regression Problems with Vector-valued Outputs

Prateek Jain, Ambuj Tewari

In regression problems involving vector-valued outputs (or equivalently, multiple responses), it is well known that the maximum likelihood estimator (MLE), which takes noise covariance structure into account, can be significantly more accurate than the ordinary least squares (OLS) estimator. However, existing literature compares OLS and MLE in terms of their asymptotic, not finite sample, guarantees. More crucially, computing the MLE in general requires solving a non-convex optimization problem and is not known to be efficiently solvable. We provide finite sample upper and lower bounds on the estimation error of OLS and MLE, in two popular models: a) Pooled model, b) Seemingly Unrelated Regression (SUR) model. We provide precise instances where the MLE is significantly more accurate than OLS. Furthermore, for both models, we show that the output of a computationally efficient alternating minimization procedure enjoys the same performance guarantee as MLE, up to universal constants. Finally, we show that for high-dimensional settings as well, the alternating minimization procedure leads to significantly more accurate solutions than the corresponding OLS solutions but with error bound that depends only logarithmically on the data dimensionality.

****

Anytime Influence Bounds and the Explosive Behavior of Continuous-Time Diffusion Networks

Kevin Scaman, Rémi Lemonnier, Nicolas Vayatis

The paper studies transition phenomena in information cascades observed along a diffusion process over some graph. We introduce the Laplace Hazard matrix and show that its spectral radius fully characterizes the dynamics of the contagion both in terms of influence and of explosion time. Using this concept, we prove tight non-asymptotic bounds for the influence of a set of nodes, and we also provide an in-depth analysis of the critical time after which the contagion becomes super-critical. Our contributions include formal definitions and tight lower bounds of critical explosion time. We illustrate the relevance of our theoretical results through several examples of information cascades used in epidemiology and viral marketing models. Finally, we provide a series of numerical experiments for various types of networks which confirm the tightness of the theoretical bounds.

****

Rapidly Mixing Gibbs Sampling for a Class of Factor Graphs Using Hierarchy Width

Christopher M. De Sa, Ce Zhang, Kunle Olukotun, Christopher Ré

Gibbs sampling on factor graphs is a widely used inference technique, which often produces good empirical results. Theoretical guarantees for its performance are weak: even for tree structured graphs, the mixing time of Gibbs may be exponential in the number of variables. To help understand the behavior of Gibbs sampling, we introduce a new (hyper)graph property, called hierarchy width. We show that under suitable conditions on the weights, bounded hierarchy width ensures polynomial mixing time. Our study of hierarchy width is in part motivated by a class of factor graph templates, hierarchical templates, which have bounded hierarchy width—regardless of the data used to instantiate them. We demonstrate a rich application from natural language processing in which Gibbs sampling provably mixes rapidly and achieves accuracy that exceeds human volunteers.

****

A Reduced-Dimension fMRI Shared Response Model

Po-Hsuan (Cameron) Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James Haxby, Peter J. Ramadge

Multi-subject fMRI data is critical for evaluating the generality and validity of findings across subjects, and its effective utilization helps improve analysis sensitivity. We develop a shared response model for aggregating multi-subject fMRI data that accounts for different functional topographies among anatomically aligned datasets. Our model demonstrates improved sensitivity in identifying a shared response for a variety of datasets and anatomical brain regions of interest. Furthermore, by removing the identified shared response, it allows improved detection of group differences. The ability to identify what is shared and what is not shared opens the model to a wide range of multi-subject fMRI studies.

****

Semi-Proximal Mirror-Prox for Nonsmooth Composite Minimization

Niao He, Zaid Harchaoui

We propose a new first-order optimization algorithm to solve high-dimensional non-smooth composite minimization problems. Typical examples of such problems have an objective that decomposes into a non-smooth empirical risk part and a non-smooth regularization penalty. The proposed algorithm, called Semi-Proximal Mirror-Prox, leverages the saddle point representation of one part of the objective while handling the other part of the objective via linear minimization over the domain. The algorithm stands in contrast with more classical proximal gradient algorithms with smoothing, which require the computation of proximal operators at each iteration and can therefore be impractical for high-dimensional problems. We establish the theoretical convergence rate of Semi-Proximal Mirror-Prox, which exhibits the optimal complexity bounds for the number of calls to linear minimization oracle. We present promising experimental results showing the interest of the approach in comparison to competing methods.

****

Subset Selection by Pareto Optimization

Chao Qian, Yang Yu, Zhi-Hua Zhou

Selecting the optimal subset from a large set of variables is a fundamental problem in various learning tasks such as feature selection, sparse regression, dict ionary learning, etc. In this paper, we propose the POSS approach which employs evolutionary Pareto optimization to find a small-sized subset with good performance. We prove that for sparse regression, POSS is able to achieve the best-so-far theoretically guaranteed approximation performance efficiently. Particularly, for the \emph{Exponential Decay} subclass, POSS is proven to achieve an optimal solution. Empirical study verifies the theoretical results, and exhibits the superior performance of POSS to greedy and convex relaxation methods.

****

Parallel Correlation Clustering on Big Graphs

Xinghao Pan, Dimitris Papailiopoulos, Samet Oymak, Benjamin Recht, Kannan Ramchandran, Michael I. Jordan

Given a similarity graph between items, correlation clustering (CC) groups similar items together and dissimilar ones apart. One of the most popular CC algorithms is KwikCluster: an algorithm that serially clusters neighborhoods of vertices, and obtains a 3-approximation ratio. Unfortunately, in practice KwikCluster requires a large number of clustering rounds, a potential bottleneck for large graphs.We present C4 and ClusterWild!, two algorithms for parallel correlation clustering that run in a polylogarithmic number of rounds, and provably achieve nearly linear speedups. C4 uses concurrency control to enforce serializability of a parallel clustering process, and guarantees a 3-approximation ratio. ClusterWild! is a coordination free algorithm that abandons consistency for the benefit of better scaling; this leads to a provably small loss in the 3 approximation ratio.We provide extensive experimental results for both algorithms, where we outperform the state of the art, both in terms of clustering accuracy and running time. We show that our algorithms can cluster billion-edge graphs in under 5 seconds on 32 cores, while achieving a 15x speedup.

****

Fast Two-Sample Testing with Analytic Representations of Probability Measures

Kacper P. Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, Arthur Gretton

We propose a class of nonparametric two-sample tests with a cost linear in the sample size. Two tests are given, both based on an ensemble of distances between analytic functions representing each of the distributions. The first test uses smoothed empirical characteristic functions to represent the distributions, the second uses distribution embeddings in a reproducing kernel Hilbert space. Analyticity implies that differences in the distributions may be detected almost surely at a finite number of randomly chosen locations/frequencies. The new tests are consistent against a larger class of alternatives than the previous linear-time tests based on the (non-smoothed) empirical characteristic functions, while being much faster than the current state-of-the-art quadratic-time kernel-based or energy distance-based tests. Experiments on artificial benchmarks and on challenging real-world testing problems demonstrate that our tests give a better power/time tradeoff than competing approaches, and in some cases, better outright power than even the most expensive quadratic-time tests. This performance advantage is retained even in high dimensions, and in cases where the difference in distributions is not observable with low order statistics.

****

A Recurrent Latent Variable Model for Sequential Data

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C. Courville, Yoshua Bengio

In this paper, we explore the inclusion of latent random variables into the hidden state of a recurrent neural network (RNN) by combining the elements of the variational autoencoder. We argue that through the use of high-level latent random variables, the variational RNN (VRNN) can model the kind of variability observed in highly structured sequential data such as natural speech. We empirically evaluate the proposed model against other related sequential models on four speech datasets and one handwriting dataset. Our results show the important roles that latent random variables can play in the RNN dynamics.

****

Unsupervised Learning by Program Synthesis

Kevin Ellis, Armando Solar-Lezama, Josh Tenenbaum

We introduce an unsupervised learning algorithmthat combines probabilistic modeling with solver-based techniques for program synthesis.We apply our techniques to both a visual learning domain and a language learning problem,showing that our algorithm can learn many visual concepts from only a few examplesand that it can recover some English inflectional morphology.Taken together, these results give both a new approach to unsupervised learning of symbolic compositional structures,and a technique for applying program synthesis tools to noisy data.

****

Learning Causal Graphs with Small Interventions

Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G. Dimakis, Sriram Vishwanath

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Learning to Transduce with Unbounded Memory

Edward Grefenstette, Karl Moritz Hermann, Mustafa Suleyman, Phil Blunsom

Recently, strong results have been demonstrated by Deep Recurrent Neural Networks on natural language transduction problems. In this paper we explore the representational power of these models using synthetic grammars designed to exhibit phenomena similar to those found in real transduction problems such as machine translation. These experiments lead us to propose new memory-based recurrent networks that implement continuously differentiable analogues of traditional data structures such as Stacks, Queues, and DeQues. We show that these architectures exhibit superior generalisation performance to Deep RNNs and are often able to learn the underlying generating algorithms in our transduction experiments.

****

Submodular Hamming Metrics

Jennifer A. Gillenwater, Rishabh K. Iyer, Bethany Lusch, Rahul Kidambi, Jeff A. Bilmes

We show that there is a largely unexplored class of functions (positive polymatroids) that can define proper discrete metrics over pairs of binary vectors and that are fairly tractable to optimize over. By exploiting submodularity, we are able to give hardness results and approximation algorithms for optimizing over such metrics. Additionally, we demonstrate empirically the effectiveness of these metrics and associated algorithms on both a metric minimization task (a form of clustering) and also a metric maximization task (generating diverse k-best lists).

****

Frank-Wolfe Bayesian Quadrature: Probabilistic Integration with Theoretical Guarantees

François-Xavier Briol, Chris Oates, Mark Girolami, Michael A. Osborne

There is renewed interest in formulating integration as an inference problem, motivated by obtaining a full distribution over numerical error that can be propagated through subsequent computation. Current methods, such as Bayesian Quadrature, demonstrate impressive empirical performance but lack theoretical analysis. An important challenge is to reconcile these probabilistic integrators with rigorous convergence guarantees. In this paper, we present the first probabilistic integrator that admits such theoretical treatment, called Frank-Wolfe Bayesian Quadrature (FWBQ). Under FWBQ, convergence to the true value of the integral is shown to be exponential and posterior contraction rates are proven to be superexponential. In simulations, FWBQ is competitive with state-of-the-art methods and out-performs alternatives based on Frank-Wolfe optimisation. Our approach is applied to successfully quantify numerical error in the solution to a challenging model choice problem in cellular biology.

****

Deep Knowledge Tracing

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J. Guibas, Jascha Sohl-Dickstein

Knowledge tracing, where a machine models the knowledge of a student as they int

eract with coursework, is an established and significantly unsolved problem in computer supported education.In this paper we explore the benefit of using recurrent neural networks to model student learning.This family of models have important advantages over current state of the art methods in that they do not require the explicit encoding of human domain knowledge,and have a far more flexible functional form which can capture substantially more complex student interactions.We show that these neural networks outperform the current state of the art in prediction on real student data,while allowing straightforward interpretation and discovery of structure in the curriculum.These results suggest a promising new line of research for knowledge tracing.
****
Generalization in Adaptive Data Analysis and Holdout Reuse
Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, Aaron Roth
Overfitting is the bane of data analysts, even when data are plentiful. Formal approaches to understanding this problem focus on statistical inference and generalization of individual analysis procedures. Yet the practice of data analysis is an inherently interactive and adaptive process: new analyses and hypotheses are proposed after seeing the results of previous ones, parameters are tuned on the basis of obtained results, and datasets are shared and reused. An investigation of this gap has recently been initiated by the authors in (Dwork et al., 2014), where we focused on the problem of estimating expectations of adaptively chosen functions.In this paper, we give a simple and practical method for reusing a holdout (or testing) set to validate the accuracy of hypotheses produced by a learning algorithm operating on a training set. Reusing a holdout set adaptively multiple times can easily lead to overfitting to the holdout set itself. We give an algorithm that enables the validation of a large number of adaptively chosen hypotheses, while provably avoiding overfitting. We illustrate the advantages of our algorithm over the standard use of the holdout set via a simple synthetic experiment.We also formalize and address the general problem of data reuse in adaptive data analysis. We show how the differential-privacy based approach in (Dwork et al., 2014) is applicable much more broadly to adaptive data analysis. We then show that a simple approach based on description length can also be used to give guarantees of statistical validity in adaptive settings. Finally, we demonstrate that these incomparable approaches can be unified via the notion of approximate max-information that we introduce. This, in particular, allows the preservation of statistical validity guarantees even when an analyst adaptively composes algorithms which have guarantees based on either of the two approaches.
****
Tractable Learning for Complex Probability Queries
Jessa Bekker, Jesse Davis, Arthur Choi, Adnan Darwiche, Guy Van den Broeck
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****
Variational Dropout and the Local Reparameterization Trick
Durk P. Kingma, Tim Salimans, Max Welling
We explore an as yet unexploited opportunity for drastically improving the efficiency of stochastic gradient variational Bayes (SGVB) with global model parameters. Regular SGVB estimators rely on sampling of parameters once per minibatch of data, and have variance that is constant w.r.t. the minibatch size. The efficiency of such estimators can be drastically improved upon by translating uncertainty about global parameters into local noise that is independent across datapoints in the minibatch. Such reparameterizations with local noise can be trivially parallelized and have variance that is inversely proportional to the minibatch size, generally leading to much faster convergence.We find an important connection with regularization by dropout: the original Gaussian dropout objective corresponds to SGVB with local noise, a scale-invariant prior and proportionally fixed posterior variance. Our method allows inference of more flexibly parameterized p

osteriors; specifically, we propose \emph{variational dropout}, a generalization of Gaussian dropout, but with a more flexibly parameterized posterior, often leading to better generalization. The method is demonstrated through several experiments.

****

Fast, Provable Algorithms for Isotonic Regression in all L_p-norms

Rasmus Kyng, Anup Rao, Sushant Sachdeva

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

On the Global Linear Convergence of Frank-Wolfe Optimization Variants

Simon Lacoste-Julien, Martin Jaggi

The Frank-Wolfe (FW) optimization algorithm has lately re-gained popularity thanks in particular to its ability to nicely handle the structured constraints appearing in machine learning applications. However, its convergence rate is known to be slow (sublinear) when the solution lies at the boundary. A simple less-known fix is to add the possibility to take away steps' during optimization, an operation that importantly does not require a feasibility oracle. In this paper, we highlight and clarify several variants of the Frank-Wolfe optimization algorithm that has been successfully applied in practice: FW with away steps, pairwise FW, fully-corrective FW and Wolfe's minimum norm point algorithm, and prove for the first time that they all enjoy global linear convergence under a weaker condition than strong convexity. The constant in the convergence rate has an elegant interpretation as the product of the (classical) condition number of the function with a novel geometric quantity that plays the role of thecondition number' of the constraint set. We provide pointers to where these algorithms have made a difference in practice, in particular with the flow polytope, the marginal polytope and the base polytope for submodular optimization.

****

Efficient Learning by Directed Acyclic Graph For Resource Constrained Prediction

Joseph Wang, Kirill Trapeznikov, Venkatesh Saligrama

We study the problem of reducing test-time acquisition costs in classification systems. Our goal is to learn decision rules that adaptively select sensors for each example as necessary to make a confident prediction. We model our system as a directed acyclic graph (DAG) where internal nodes correspond to sensor subsets and decision functions at each node choose whether to acquire a new sensor or classify using the available measurements. This problem can be naturally posed as an empirical risk minimization over training data. Rather than jointly optimizing such a highly coupled and non-convex problem over all decision nodes, we propose an efficient algorithm motivated by dynamic programming. We learn node policies in the DAG by reducing the global objective to a series of cost sensitive learning problems. Our approach is computationally efficient and has proven guarantees of convergence to the optimal system for a fixed architecture. In addition, we present an extension to map other budgeted learning problems with large number of sensors to our DAG architecture and demonstrate empirical performance exceeding state-of-the-art algorithms for data composed of both few and many sensors.

****

Finite-Time Analysis of Projected Langevin Monte Carlo

Sebastien Bubeck, Ronen Eldan, Joseph Lehec

We analyze the projected Langevin Monte Carlo (LMC) algorithm, a close cousin of projected Stochastic Gradient Descent (SGD). We show that LMC allows to sample in polynomial time from a posterior distribution restricted to a convex body and with concave log-likelihood. This gives the first Markov chain to sample from a log-concave distribution with a first-order oracle, as the existing chains with provable guarantees (lattice walk, ball walk and hit-and-run) require a zeroth-order oracle. Our proof uses elementary concepts from stochastic calculus which could be useful more generally to understand SGD and its variants.

****

A Universal Catalyst for First-Order Optimization

Hongzhou Lin, Julien Mairal, Zaid Harchaoui

We introduce a generic scheme for accelerating first-order optimization methods in the sense of Nesterov, which builds upon a new analysis of the accelerated proximal point algorithm. Our approach consists of minimizing a convex objective by approximately solving a sequence of well-chosen auxiliary problems, leading to faster convergence. This strategy applies to a large class of algorithms, including gradient descent, block coordinate descent, SAG, SAGA, SDCA, SVRG, Finito/MISO, and their proximal variants. For all of these methods, we provide acceleration and explicit support for non-strongly convex objectives. In addition to theoretical speed-up, we also show that acceleration is useful in practice, especially for ill-conditioned problems where we measure significant improvements.

****

Distributed Submodular Cover: Succinctly Summarizing Massive Data

Baharan Mirzasoleiman, Amin Karbasi, Ashwinkumar Badanidiyuru, Andreas Krause

How can one find a subset, ideally as small as possible, that well represents a massive dataset? I.e., its corresponding utility, measured according to a suitable utility function, should be comparable to that of the whole dataset. In this paper, we formalize this challenge as a submodular cover problem. Here, the utility is assumed to exhibit submodularity, a natural diminishing returns condition preva- lent in many data summarization applications. The classical greedy algorithm is known to provide solutions with logarithmic approximation guarantees compared to the optimum solution. However, this sequential, centralized approach is imprac- tical for truly large-scale problems. In this work, we develop the first distributed algorithm – DISCOVER – for submodular set cover that is easily implementable using MapReduce-style computations. We theoretically analyze our approach, and present approximation guarantees for the solutions returned by DISCOVER. We also study a natural trade-off between the communication cost and the num- ber of rounds required to obtain such a solution. In our extensive experiments, we demonstrate the effectiveness of our approach on several applications, includ- ing active set selection, exemplar based clustering, and vertex cover on tens of millions of data points using Spark.

****

Robust PCA with compressed data

Wooseok Ha, Rina Foygel Barber

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Bidirectional Recurrent Convolutional Networks for Multi-Frame Super-Resolution

Yan Huang, Wei Wang, Liang Wang

Super resolving a low-resolution video is usually handled by either single-image super-resolution (SR) or multi-frame SR. Single-Image SR deals with each video frame independently, and ignores intrinsic temporal dependency of video frames which actually plays a very important role in video super-resolution. Multi-Frame SR generally extracts motion information, e.g. optical flow, to model the temporal dependency, which often shows high computational cost. Considering that recurrent neural network (RNN) can model long-term contextual information of temporal sequences well, we propose a bidirectional recurrent convolutional network for efficient multi-frame SR.Different from vanilla RNN, 1) the commonly-used recurrent full connections are replaced with weight-sharing convolutional connections and 2) conditional convolutional connections from previous input layers to current hidden layer are added for enhancing visual-temporal dependency modelling. With the powerful temporal dependency modelling, our model can super resolve videos with complex motions and achieve state-of-the-art performance. Due to the cheap convolution operations, our model has a low computational complexity and runs orders of magnitude faster than other multi-frame methods.

****

Regret-Based Pruning in Extensive-Form Games

Noam Brown, Tuomas Sandholm

Counterfactual Regret Minimization (CFR) is a leading algorithm for finding a Nash equilibrium in large zero-sum imperfect-information games. CFR is an iterative algorithm that repeatedly traverses the game tree, updating regrets at each information set.We introduce an improvement to CFR that prunes any path of play in the tree, and its descendants, that has negative regret. It revisits that sequence at the earliest subsequent CFR iteration where the regret could have become positive, had that path been explored on every iteration. The new algorithm maintains CFR's convergence guarantees while making iterations significantly faster---even if previously known pruning techniques are used in the comparison. This improvement carries over to CFR+, a recent variant of CFR. Experiments show an order of magnitude speed improvement, and the relative speed improvement increases with the size of the game.

****

Adaptive Low-Complexity Sequential Inference for Dirichlet Process Mixture Models

Theodoros Tsiligkaridis, Theodoros Tsiligkaridis, Keith Forsythe

We develop a sequential low-complexity inference procedure for Dirichlet process mixtures of Gaussians for online clustering and parameter estimation when the number of clusters are unknown a-priori. We present an easily computable, closed form parametric expression for the conditional likelihood, in which hyperparameters are recursively updated as a function of the streaming data assuming conjugate priors. Motivated by large-sample asymptotics, we propose a noveladaptive low-complexity design for the Dirichlet process concentration parameter and show that the number of classes grow at most at a logarithmic rate. We further prove that in the large-sample limit, the conditional likelihood and datapredictive distribution become asymptotically Gaussian. We demonstrate through experiments on synthetic and real data sets that our approach is superior to otheronline state-of-the-art methods.

****

Bidirectional Recurrent Neural Networks as Generative Models

Mathias Berglund, Tapani Raiko, Mikko Honkala, Leo Kärkkäinen, Akos Vetek, Juha T. Karhunen

Bidirectional recurrent neural networks (RNN) are trained to predict both in the positive and negative time directions simultaneously. They have not been used commonly in unsupervised tasks, because a probabilistic interpretation of the model has been difficult. Recently, two different frameworks, GSN and NADE, provide a connection between reconstruction and probabilistic modeling, which makes the interpretation possible. As far as we know, neither GSN or NADE have been studied in the context of time series before.As an example of an unsupervised task, we study the problem of filling in gaps in high-dimensional time series with complex dynamics. Although unidirectional RNNs have recently been trained successfully to model such time series, inference in the negative time direction is non-trivial. We propose two probabilistic interpretations of bidirectional RNNs that can be used to reconstruct missing gaps efficiently. Our experiments on text data show that both proposed methods are much more accurate than unidirectional reconstructions, although a bit less accurate than a computationally complex bidirectional Bayesian inference on the unidirectional RNN. We also provide results on music data for which the Bayesian inference is computationally infeasible, demonstrating the scalability of the proposed methods.

****

Double or Nothing: Multiplicative Incentive Mechanisms for Crowdsourcing

Nihar Bhadresh Shah, Dengyong Zhou

Crowdsourcing has gained immense popularity in machine learning applications for obtaining large amounts of labeled data. Crowdsourcing is cheap and fast, but suffers from the problem of low-quality data. To address this fundamental challenge in crowdsourcing, we propose a simple payment mechanism to incentivize workers to answer only the questions that they are sure of and skip the rest. We show that surprisingly, under a mild and natural no-free-lunch requirement, this mech anism is the one and only incentive-compatible payment mechanism possible. We also show that among all possible incentive-compatible mechanisms (that may or may not satisfy no-free-lunch), our mechanism makes the smallest possible payment to spammers. Interestingly, this unique mechanism takes a multiplicative form. The simplicity of the mechanism is an added benefit. In preliminary experiments involving over several hundred workers, we observe a significant reduction in the error rates under our unique mechanism for the same or lower monetary expenditure.

****

Asynchronous stochastic convex optimization: the noise is in the noise and SGD don't care

Sorathan Chaturapruek, John C. Duchi, Christopher Ré

We show that asymptotically, completely asynchronous stochastic gradient procedures achieve optimal (even to constant factors) convergence rates for the solution of convex optimization problems under nearly the same conditions required for asymptotic optimality of standard stochastic gradient procedures. Roughly, the noise inherent to the stochastic approximation scheme dominates any noise from asynchrony. We also give empirical evidence demonstrating the strong performance of asynchronous, parallel stochastic optimization schemes, demonstrating that the robustness inherent to stochastic approximation problems allows substantially faster parallel and asynchronous solution methods. In short, we show that for many stochastic approximation problems, as Freddie Mercury sings in Queen's \emph{Bohemian Rhapsody}, ``Nothing really matters.''

****

Bounding errors of Expectation-Propagation

Guillaume P. Dehaene, Simon Barthelmé

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

On the Limitation of Spectral Methods: From the Gaussian Hidden Clique Problem to Rank-One Perturbations of Gaussian Tensors

Andrea Montanari, Daniel Reichman, Ofer Zeitouni

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Convergence Rates of Active Learning for Maximum Likelihood Estimation

Kamalika Chaudhuri, Sham M. Kakade, Praneeth Netrapalli, Sujay Sanghavi

An active learner is given a class of models, a large set of unlabeled examples, and the ability to interactively query labels of a subset of these examples; the goal of the learner is to learn a model in the class that fits the data well. Previous theoretical work has rigorously characterized label complexity of active learning, but most of this work has focused on the PAC or the agnostic PAC model. In this paper, we shift our attention to a more general setting -- maximum likelihood estimation. Provided certain conditions hold on the model class, we provide a two-stage active learning algorithm for this problem. The conditions we require are fairly general, and cover the widely popular class of Generalized Linear Models, which in turn, include models for binary and multi-class classification, regression, and conditional random fields. We provide an upper bound on the label requirement of our algorithm, and a lower bound that matches it up to lower order terms. Our analysis shows that unlike binary classification in the realizable case, just a single extraround of interaction is sufficient to achieve near-optimal performance in maximum likelihood estimation. On the empirical side, the recent work in (Gu et al. 2012) and (Gu et al. 2014) (on active linear and logistic regression) shows the promise of this approach.

****

Distributionally Robust Logistic Regression

Soroosh Shafieezadeh Abadeh, Peyman Mohajerin Mohajerin Esfahani, Daniel Kuhn

This paper proposes a distributionally robust approach to logistic regression. We use the Wasserstein distance to construct a ball in the space of probability distributions centered at the uniform distribution on the training samples. If the radius of this Wasserstein ball is chosen judiciously, we can guarantee that it contains the unknown data-generating distribution with high confidence. We then formulate a distributionally robust logistic regression model that minimizes a worst-case expected logloss function, where the worst case is taken over all distributions in the Wasserstein ball. We prove that this optimization problem admits a tractable reformulation and encapsulates the classical as well as the popular regularized logistic regression problems as special cases. We further propose a distributionally robust approach based on Wasserstein balls to compute upper and lower confidence bounds on the misclassification probability of the resulting classifier. These bounds are given by the optimal values of two highly tractable linear programs. We validate our theoretical out-of-sample guarantees through simulated and empirical experiments.

****

Adaptive Primal-Dual Splitting Methods for Statistical Learning and Image Processing

Tom Goldstein, Min Li, Xiaoming Yuan

The alternating direction method of multipliers (ADMM) is an important tool for solving complex optimization problems, but it involves minimization sub-steps that are often difficult to solve efficiently.   The Primal-Dual Hybrid Gradient (PDHG) method is a powerful alternative that often has simpler substeps than ADMM, thus producing lower complexity solvers. Despite the flexibility of this method, PDHG is often impractical because it requires the careful choice of multiple stepsize parameters. There is often no intuitive way to choose these parameters to maximize efficiency, or even achieve convergence. We propose self-adaptive stepsize rules that automatically tune PDHG parameters for optimal convergence. We rigorously analyze our methods, and identify convergence rates. Numerical experiments show that adaptive PDHG has strong advantages over non-adaptive methods in terms of both efficiency and simplicity for the user.

****

Regret Lower Bound and Optimal Algorithm in Finite Stochastic Partial Monitoring

Junpei Komiyama, Junya Honda, Hiroshi Nakagawa

Partial monitoring is a general model for sequential learning with limited feedback formalized as a game between two players. In this game, the learner chooses an action and at the same time the opponent chooses an outcome, then the learner suffers a loss and receives a feedback signal. The goal of the learner is to minimize the total loss. In this paper, we study partial monitoring with finite actions and stochastic outcomes. We derive a logarithmic distribution-dependent regret lower bound that defines the hardness of the problem. Inspired by the DMED algorithm (Honda and Takemura, 2010) for the multi-armed bandit problem, we propose PM-DMED, an algorithm that minimizes the distribution-dependent regret. PM-DMED significantly outperforms state-of-the-art algorithms in numerical experiments. To show the optimality of PM-DMED with respect to the regret bound, we slightly modify the algorithm by introducing a hinge function (PM-DMED-Hinge). Then, we derive an asymptotical optimal regret upper bound of PM-DMED-Hinge that matches the lower bound.

****

Online Prediction at the Limit of Zero Temperature

Mark Herbster, Stephen Pasteris, Shaona Ghosh

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Scalable Semi-Supervised Aggregation of Classifiers

Akshay Balsubramani, Yoav Freund

We present and empirically evaluate an efficient algorithm that learns to aggreg

ate the predictions of an ensemble of binary classifiers. The algorithm uses the structure of the ensemble predictions on unlabeled data to yield significant performance improvements. It does this without making assumptions on the structure or origin of the ensemble, without parameters, and as scalably as linear learning. We empirically demonstrate these performance gains with random forests.
****

Deep Convolutional Inverse Graphics Network

Tejas D. Kulkarni, William F. Whitney, Pushmeet Kohli, Josh Tenenbaum

This paper presents the Deep Convolution Inverse Graphics Network (DC-IGN), a model that aims to learn an interpretable representation of images, disentangled with respect to three-dimensional scene structure and viewing transformations such as depth rotations and lighting variations. The DC-IGN model is composed of multiple layers of convolution and de-convolution operators and is trained using the Stochastic Gradient Variational Bayes (SGVB) algorithm. We propose a training procedure to encourage neurons in the graphics code layer to represent a specific transformation (e.g. pose or light). Given a single input image, our model can generate new images of the same object with variations in pose and lighting. We present qualitative and quantitative tests of the model's efficacy at learning a 3D rendering engine for varied object classes including faces and chairs.
****

Smooth and Strong: MAP Inference with Linear Convergence

Ofer Meshi, Mehrdad Mahdavi, Alex Schwing

Maximum a-posteriori (MAP) inference is an important task for many applications. Although the standard formulation gives rise to a hard combinatorial optimization problem, several effective approximations have been proposed and studied in recent years. We focus on linear programming (LP) relaxations, which have achieved state-of-the-art performance in many applications. However, optimization of the resulting program is in general challenging due to non-smoothness and complex non-separable constraints.Therefore, in this work we study the benefits of augmenting the objective function of the relaxation with strong convexity. Specifically, we introduce strong convexity by adding a quadratic term to the LP relaxation objective. We provide theoretical guarantees for the resulting programs, bounding the difference between their optimal value and the original optimum. Further, we propose suitable optimization algorithms and analyze their convergence.
****

Efficient Compressive Phase Retrieval with Constrained Sensing Vectors

Sohail Bahmani, Justin Romberg

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
****

Convolutional spike-triggered covariance analysis for neural subunit models

Anqi Wu, Il Memming Park, Jonathan W. Pillow

Subunit models provide a powerful yet parsimonious description of neural spike responses to complex stimuli. They can be expressed by a cascade of two linear-nonlinear (LN) stages, with the first linear stage defined by convolution with one or more filters. Recent interest in such models has surged due to their biological plausibility and accuracy for characterizing early sensory responses. However, fitting subunit models poses a difficult computational challenge due to the expense of evaluating the log-likelihood and the ubiquity of local optima. Here we address this problem by forging a theoretical connection between spike-triggered covariance analysis and nonlinear subunit models. Specifically, we show that a ''convolutional'' decomposition of the spike-triggered average (STA) and covariance (STC) provides an asymptotically efficient estimator for the subunit model under certain technical conditions. We also prove the identifiability of such convolutional decomposition under mild assumptions. Our moment-based methods outperform highly regularized versions of the GQM on neural data from macaque primary visual cortex, and achieves nearly the same prediction performance as the full maximum-likelihood estimator, yet with substantially lowe r cost.

****

Recovering Communities in the General Stochastic Block Model Without Knowing the Parameters

Emmanuel Abbe, Colin Sandon

The stochastic block model (SBM) has recently gathered significant attention due to new threshold phenomena. However, most developments rely on the knowledge of the model parameters, or at least on the number of communities. This paper introduces efficient algorithms that do not require such knowledge and yet achieve the optimal information-theoretic tradeoffs identified in Abbe-Sandon FOCS15. In the constant degree regime, an algorithm is developed that requires only a lower-bound on the relative sizes of the communities and achieves the optimal accuracy scaling for large degrees. This lower-bound requirement is removed for the regime of arbitrarily slowly diverging degrees, and the model parameters are learned efficiently. For the logarithmic degree regime, this is further enhanced into a fully agnostic algorithm that achieves the CH-limit for exact recovery in quasi-linear time. These provide the first algorithms affording efficiency, universality and information-theoretic optimality for strong and weak consistency in the SBM.

****

On Variance Reduction in Stochastic Gradient Descent and its Asynchronous Variants

Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, Alexander J. Smola

We study optimization algorithms based on variance reduction for stochastic gradientdescent (SGD). Remarkable recent progress has been made in this directionthrough development of algorithms like SAG, SVRG, SAGA. These algorithmshave been shown to outperform SGD, both theoretically and empirically. However,asynchronous versions of these algorithms—a crucial requirement for modernlarge-scale applications—have not been studied. We bridge this gap by presentinga unifying framework that captures many variance reduction techniques.Subsequently, we propose an asynchronous algorithm grounded in our framework,with fast convergence rates. An important consequence of our general approachis that it yields asynchronous versions of variance reduction algorithms such asSVRG, SAGA as a byproduct. Our method achieves near linear speedup in sparsesettings common to machine learning. We demonstrate the empirical performanceof our method through a concrete realization of asynchronous SVRG.

****

Galileo: Perceiving Physical Object Properties by Integrating a Physics Engine with Deep Learning

Jiajun Wu, Ilker Yildirim, Joseph J. Lim, Bill Freeman, Josh Tenenbaum

Humans demonstrate remarkable abilities to predict physical events in dynamic scenes, and to infer the physical properties of objects from static images. We propose a generative model for solving these problems of physical scene understanding from real-world videos and images. At the core of our generative model is a 3D physics engine, operating on an object-based representation of physical properties, including mass, position, 3D shape, and friction. We can infer these latent properties using relatively brief runs of MCMC, which drive simulations in the physics engine to fit key features of visual observations. We further explore directly mapping visual inputs to physical properties, inverting a part of the generative process using deep learning. We name our model Galileo, and evaluate it on a video dataset with simple yet physically rich scenarios. Results show that Galileo is able to infer the physical properties of objects and predict the outcome of a variety of physical events, with an accuracy comparable to human subjects. Our study points towards an account of human vision with generative physical knowledge at its core, and various recognition models as helpers leading to efficient inference.

****

Optimal Rates for Random Fourier Features

Bharath Sriperumbudur, Zoltan Szabo

Kernel methods represent one of the most powerful tools in machine learning to t ackle problems expressed in terms of function values and derivatives due to their capability to represent and model complex relations. While these methods show good versatility, they are computationally intensive and have poor scalability to large data as they require operations on Gram matrices. In order to mitigate this serious computational limitation, recently randomized constructions have been proposed in the literature, which allow the application of fast linear algorithms. Random Fourier features (RFF) are among the most popular and widely applied constructions: they provide an easily computable, low-dimensional feature representation for shift-invariant kernels. Despite the popularity of RFFs, very little is understood theoretically about their approximation quality. In this paper, we provide a detailed finite-sample theoretical analysis about the approximation quality of RFFs by (i) establishing optimal (in terms of the RFF dimension, and growing set size) performance guarantees in uniform norm, and (ii) presenting guarantees in L^r (1 ≤ r < ∞) norms. We also propose an RFF approximation to derivatives of a kernel with a theoretical study on its approximation quality.

****

Deep learning with Elastic Averaging SGD

Sixin Zhang, Anna E. Choromanska, Yann LeCun

We study the problem of stochastic optimization for deep learning in the parallel computing environment under communication constraints. A new algorithm is proposed in this setting where the communication and coordination of work among concurrent processes (local workers), is based on an elastic force which links the parameters they compute with a center variable stored by the parameter server (master). The algorithm enables the local workers to perform more exploration, i.e. the algorithm allows the local variables to fluctuate further from the center variable by reducing the amount of communication between local workers and the master. We empirically demonstrate that in the deep learning setting, due to the existence of many local optima, allowing more exploration can lead to the improved performance. We propose synchronous and asynchronous variants of the new algorithm. We provide the stability analysis of the asynchronous variant in the round-robin scheme and compare it with the more common parallelized method ADMM. We show that the stability of EASGD is guaranteed when a simple stability condition is satisfied, which is not the case for ADMM. We additionally propose the momentum-based version of our algorithm that can be applied in both synchronous and asynchronous settings. Asynchronous variant of the algorithm is applied to train convolutional neural networks for image classification on the CIFAR and ImageNet datasets. Experiments demonstrate that the new algorithm accelerates the training of deep architectures compared to DOWNPOUR and other common baseline approaches and furthermore is very communication efficient.

****

Online F-Measure Optimization

Róbert Busa-Fekete, Balázs Szörényi, Krzysztof Dembczynski, Eyke Hüllermeier

The F-measure is an important and commonly used performance metric for binary prediction tasks. By combining precision and recall into a single score, it avoids disadvantages of simple metrics like the error rate, especially in cases of imbalanced class distributions. The problem of optimizing the F-measure, that is, of developing learning algorithms that perform optimally in the sense of this measure, has recently been tackled by several authors. In this paper, we study the problem of F-measure maximization in the setting of online learning. We propose an efficient online algorithm and provide a formal analysis of its convergence properties. Moreover, first experimental results are presented, showing that our method performs well in practice.

****

Bayesian Manifold Learning: The Locally Linear Latent Variable Model (LL-LVM)

Mijung Park, Wittawat Jitkrittum, Ahmad Qamar, Zoltan Szabo, Lars Buesing, Maneesh Sahani

We introduce the Locally Linear Latent Variable Model (LL-LVM), a probabilistic model for non-linear manifold discovery that describes a joint distribution over observations, their manifold coordinates and locally linear maps conditioned on a set of neighbourhood relationships. The model allows straightforward variatio nal optimisation of the posterior distribution on coordinates and locally linear maps from the latent space to the observation space given the data. Thus, the LL-LVM encapsulates the local-geometry preserving intuitions that underlie non-probabilistic methods such as locally linear embedding (LLE). Its probabilistic semantics make it easy to evaluate the quality of hypothesised neighbourhood relationships, select the intrinsic dimensionality of the manifold, construct out-of-sample extensions and to combine the manifold model with additional probabilistic models that capture the structure of coordinates within the manifold.

****

Smooth Interactive Submodular Set Cover

Bryan D. He, Yisong Yue

Interactive submodular set cover is an interactive variant of submodular set cover over a hypothesis class of submodular functions, where the goal is to satisfy all sufficiently plausible submodular functions to a target threshold using as few (cost-weighted) actions as possible. It models settings where there is uncertainty regarding which submodular function to optimize. In this paper, we propose a new extension, which we call smooth interactive submodular set cover, that allows the target threshold to vary depending on the plausibility of each hypothesis. We present the first algorithm for this more general setting with theoretical guarantees on optimality. We further show how to extend our approach to deal with real-valued functions, which yields new theoretical results for real-valued submodular set cover for both the interactive and non-interactive settings.

****

Column Selection via Adaptive Sampling

Saurabh Paul, Malik Magdon-Ismail, Petros Drineas

Selecting a good column (or row) subset of massive data matrices has found many applications in data analysis and machine learning. We propose a new adaptive sampling algorithm that can be used to improve any relative-error column selection algorithm. Our algorithm delivers a tighter theoretical bound on the approximation error which we also demonstrate empirically using two well known relative-error column subset selection algorithms. Our experimental results on synthetic and real-world data show that our algorithm outperforms non-adaptive sampling as well as prior adaptive sampling approaches.

****

Parallel Multi-Dimensional LSTM, With Application to Fast Biomedical Volumetric Image Segmentation

Marijn F. Stollenga, Wonmin Byeon, Marcus Liwicki, Jürgen Schmidhuber

Convolutional Neural Networks (CNNs) can be shifted across 2D images or 3D videos to segment them. They have a fixed input size and typically perceive only small local contexts of the pixels to be classified as foreground or background. In contrast, Multi-Dimensional Recurrent NNs (MD-RNNs) can perceive the entire spatio-temporal context of each pixel in a few sweeps through all pixels, especially when the RNN is a Long Short-Term Memory (LSTM). Despite these theoretical advantages, however, unlike CNNs, previous MD-LSTM variants were hard to parallelise on GPUs. Here we re-arrange the traditional cuboid order of computations in MD-LSTM in pyramidal fashion. The resulting PyraMiD-LSTM is easy to parallelise, especially for 3D data such as stacks of brain slice images. PyraMiD-LSTM achieved best known pixel-wise brain image segmentation results on MRBrainS13 (and competitive results on EM-ISBI12).

****

Discriminative Robust Transformation Learning

Jiaji Huang, Qiang Qiu, Guillermo Sapiro, Robert Calderbank

This paper proposes a framework for learning features that are robust to data variation, which is particularly important when only a limited number of training samples are available. The framework makes it possible to tradeoff the discriminative value of learned features against the generalization error of the learning algorithm. Robustness is achieved by encouraging the transform that maps data to features to be a local isometry. This geometric property is shown to improve (K, \epsilon)-robustness, thereby providing theoretical justification for reductions in generalization error observed in experiments. The proposed optimization fr ameworkis used to train standard learning algorithms such as deep neural networks. Experimental results obtained on benchmark datasets, such as labeled faces in the wild,demonstrate the value of being able to balance discrimination and robustness.

****

Deep Poisson Factor Modeling

Ricardo Henao, Zhe Gan, James Lu, Lawrence Carin

We propose a new deep architecture for topic modeling, based on Poisson Factor Analysis (PFA) modules. The model is composed of a Poisson distribution to model observed vectors of counts, as well as a deep hierarchy of hidden binary units. Rather than using logistic functions to characterize the probability that a latent binary unit is on, we employ a Bernoulli-Poisson link, which allows PFA modules to be used repeatedly in the deep architecture. We also describe an approach to build discriminative topic models, by adapting PFA modules. We derive efficient inference via MCMC and stochastic variational methods, that scale with the number of non-zeros in the data and binary units, yielding significant efficiency, relative to models based on logistic links. Experiments on several corpora demonstrate the advantages of our model when compared to related deep models.

****

Max-Margin Majority Voting for Learning from Crowds

TIAN TIAN, Jun Zhu

Learning-from-crowds aims to design proper aggregation strategies to infer the unknown true labels from the noisy labels provided by ordinary web workers. This paper presents max-margin majority voting (M^3V) to improve the discriminative ability of majority voting and further presents a Bayesian generalization to incorporate the flexibility of generative methods on modeling noisy observations with worker confusion matrices. We formulate the joint learning as a regularized Bayesian inference problem, where the posterior regularization is derived by maximizing the margin between the aggregated score of a potential true label and that of any alternative label. Our Bayesian model naturally covers the Dawid-Skene estimator and M^3V. Empirical results demonstrate that our methods are competitive, often achieving better results than state-of-the-art estimators.

****

Competitive Distribution Estimation: Why is Good-Turing Good

Alon Orlitsky, Ananda Theertha Suresh

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Embedding Inference for Structured Multilabel Prediction

Farzaneh Mirzazadeh, Siamak Ravanbakhsh, Nan Ding, Dale Schuurmans

A key bottleneck in structured output prediction is the need for inference during training and testing, usually requiring some form of dynamic programming. Rather than using approximate inference or tailoring a specialized inference method for a particular structure---standard responses to the scaling challenge---we propose to embed prediction constraints directly into the learned representation. By eliminating the need for explicit inference a more scalable approach to structured output prediction can be achieved, particularly at test time. We demonstrate the idea for multi-label prediction under subsumption and mutual exclusion constraints, where a relationship to maximum margin structured output prediction can be established. Experiments demonstrate that the benefits of structured output training can still be realized even after inference has been eliminated.

****

Spectral Learning of Large Structured HMMs for Comparative Epigenomics

Chicheng Zhang, Jimin Song, Kamalika Chaudhuri, Kevin Chen

We develop a latent variable model and an efficient spectral algorithm motivated by the recent emergence of very large data sets of chromatin marks from multiple human cell types. A natural model for chromatin data in one cell type is a Hidden Markov Model (HMM); we model the relationship between multiple cell types by connecting their hidden states by a fixed tree of known structure. The main challenge with learning parameters of such models is that iterative methods such as EM are very slow, while naive spectral methods result in time and space complexity exponential in the number of cell types. We exploit properties of the tree structure of the hidden states to provide spectral algorithms that are more computationally efficient for current biological datasets. We provide sample complexity bounds for our algorithm and evaluate it experimentally on biological data from nine human cell types. Finally, we show that beyond our specific model, some of our algorithmic ideas can be applied to other graphical models.

****

Deeply Learning the Messages in Message Passing Inference

Guosheng Lin, Chunhua Shen, Ian Reid, Anton van den Hengel

Deep structured output learning shows great promise in tasks like semantic image segmentation. We proffer a new, efficient deep structured model learning scheme, in which we show how deep Convolutional Neural Networks (CNNs) can be used to directly estimate the messages in message passing inference for structured prediction with Conditional Random Fields CRFs). With such CNN message estimators, we obviate the need to learn or evaluate potential functions for message calculation. This confers significant efficiency for learning, since otherwise when performing structured learning for a CRF with CNN potentials it is necessary to undertake expensive inference for every stochastic gradient iteration. The network output dimension of message estimators is the same as the number of classes, rather than exponentially growing in the order of the potentials. Hence it is more scalable for cases that a large number of classes are involved. We apply our method to semantic image segmentation and achieve impressive performance, which demonstrates the effectiveness and usefulness of our CNN message learning method.

****

Bayesian Active Model Selection with an Application to Automated Audiometry

Jacob Gardner, Gustavo Malkomes, Roman Garnett, Kilian Q. Weinberger, Dennis Barbour, John P. Cunningham

We introduce a novel information-theoretic approach for active model selection and demonstrate its effectiveness in a real-world application. Although our method can work with arbitrary models, we focus on actively learning the appropriate structure for Gaussian process (GP) models with arbitrary observation likelihoods. We then apply this framework to rapid screening for noise-induced hearing loss (NIHL), a widespread and preventible disability, if diagnosed early. We construct a GP model for pure-tone audiometric responses of patients with NIHL. Using this and a previously published model for healthy responses, the proposed method is shown to be capable of diagnosing the presence or absence of NIHL with drastically fewer samples than existing approaches. Further, the method is extremely fast and enables the diagnosis to be performed in real time.

****

Collaboratively Learning Preferences from Ordinal Data

Sewoong Oh, Kiran K. Thekumparampil, Jiaming Xu

In personalized recommendation systems, it is important to predict preferences of a user on items that have not been seen by that user yet. Similarly, in revenue management, it is important to predict outcomes of comparisons among those items that have never been compared so far. The MultiNomial Logit model, a popular discrete choice model, captures the structure of the hidden preferences with a low-rank matrix. In order to predict the preferences, we want to learn the underlying model from noisy observations of the low-rank matrix, collected as revealed preferences in various forms of ordinal data. A natural approach to learn such a model is to solve a convex relaxation of nuclear norm minimization. We present the convex relaxation approach in two contexts of interest: collaborative ranking and bundled choice modeling. In both cases, we show that the convex relaxation is minimax optimal. We prove an upper bound on the resulting error with finite samples, and provide a matching information-theoretic lower bound.

****

Shepard Convolutional Neural Networks

Jimmy SJ Ren, Li Xu, Qiong Yan, Wenxiu Sun

Deep learning has recently been introduced to the field of low-level computer vision and image processing. Promising results have been obtained in a number of tasks including super-resolution, inpainting, deconvolution, filtering, etc. However, previously adopted neural network approaches such as convolutional neural networks and sparse auto-encoders are inherently with translation invariant operators. We found this property prevents the deep learning approaches from outperforming the state-of-the-art if the task itself requires translation variant interpolation (TVI). In this paper, we draw on Shepard interpolation and design Shepard Convolutional Neural Networks (ShCNN) which efficiently realizes end-to-end trainable TVI operators in the network. We show that by adding only a few feature maps in the new Shepard layers, the network is able to achieve stronger results than a much deeper architecture. Superior performance on both image inpainting and super-resolution is obtained where our system outperforms previous ones while keeping the running time competitive.

****

Learning Wake-Sleep Recurrent Attention Models

Jimmy Ba, Russ R. Salakhutdinov, Roger B. Grosse, Brendan J. Frey

Despite their success, convolutional neural networks are computationally expensive because they must examine all image locations. Stochastic attention-based models have been shown to improve computational efficiency at test time, but they remain difficult to train because of intractable posterior inference and high variance in the stochastic gradient estimates. Borrowing techniques from the literature on training deep generative models, we present the Wake-Sleep Recurrent Attention Model, a method for training stochastic attention networks which improves posterior inference and which reduces the variability in the stochastic gradients. We show that our method can greatly speed up the training time for stochastic attention networks in the domains of image classification and caption generation.

****

Matrix Manifold Optimization for Gaussian Mixtures

Reshad Hosseini, Suvrit Sra

We take a new look at parameter estimation for Gaussian Mixture Model (GMMs). Specifically, we advance Riemannian manifold optimization (on the manifold of positive definite matrices) as a potential replacement for Expectation Maximization (EM), which has been the de facto standard for decades. An out-of-the-box invocation of Riemannian optimization, however, fails spectacularly: it obtains the same solution as EM, but vastly slower. Building on intuition from geometric convexity, we propose a simple reformulation that has remarkable consequences: it makes Riemannian optimization not only match EM (a nontrivial result on its own, given the poor record nonlinear programming has had against EM), but also outperform it in many settings. To bring our ideas to fruition, we develop a well-tuned Riemannian LBFGS method that proves superior to known competing methods (e.g., Riemannian conjugate gradient). We hope that our results encourage a wider consideration of manifold optimization in machine learning and statistics.

****

Minimum Weight Perfect Matching via Blossom Belief Propagation

Sung-Soo Ahn, Sejun Park, Michael Chertkov, Jinwoo Shin

Max-product Belief Propagation (BP) is a popular message-passing algorithm for computing a Maximum-A-Posteriori (MAP) assignment over a distribution represented by a Graphical Model (GM). It has been shown that BP can solve a number of combinatorial optimization problems including minimum weight matching, shortest path, network flow and vertex cover under the following common assumption: the respective Linear Programming (LP) relaxation is tight, i.e., no integrality gap is present. However, when LP shows an integrality gap, no model has been known which can be solved systematically via sequential applications of BP. In this paper, we develop the first such algorithm, coined Blossom-BP, for solving the minimum weight matching problem over arbitrary graphs. Each step of the sequential algorithm requires applying BP over a modified graph constructed by contractions and expansions of blossoms, i.e., odd sets of vertices. Our scheme guarantees termination in O(n^2) of BP runs, where n is the number of vertices in the original graph. In essence, the Blossom-BP offers a distributed version of the celebrated Edmonds' Blossom algorithm by jumping at once over many sub-steps with a single BP. Moreover, our result provides an interpretation of the Edmonds' algorithm as a sequence of LPs.

****

Human Memory Search as Initial-Visit Emitting Random Walk

Kwang-Sung Jun, Jerry Zhu, Timothy T. Rogers, Zhuoran Yang, ming yuan

Imagine a random walk that outputs a state only when visiting it for the first time. The observed output is therefore a repeat-censored version of the underlying walk, and consists of a permutation of the states or a prefix of it. We call this model initial-visit emitting random walk (INVITE). Prior work has shown that the random walks with such a repeat-censoring mechanism explain well human behavior in memory search tasks, which is of great interest in both the study of human cognition and various clinical applications. However, parameter estimation in INVITE is challenging, because naive likelihood computation by marginalizing over infinitely many hidden random walk trajectories is intractable. In this paper, we propose the first efficient maximum likelihood estimate (MLE) for INVITE by decomposing the censored output into a series of absorbing random walks. We also prove theoretical properties of the MLE including identifiability and consistency. We show that INVITE outperforms several existing methods on real-world human response data from memory search tasks.

****

Bounding the Cost of Search-Based Lifted Inference

David B. Smith, Vibhav G. Gogate

Recently, there has been growing interest in systematic search-based and importance sampling-based lifted inference algorithms for statistical relational models (SRMs). These lifted algorithms achieve significant complexity reductions over their propositional counterparts by using lifting rules that leverage symmetries in the relational representation. One drawback of these algorithms is that they use an inference-blind representation of the search space, which makes it difficult to efficiently pre-compute tight upper bounds on the exact cost of inference without running the algorithm to completion. In this paper, we present a principled approach to address this problem. We introduce a lifted analogue of the propositional And/Or search space framework, which we call a lifted And/Or schematic. Given a schematic-based representation of an SRM, we show how to efficiently compute a tight upper bound on the time and space cost of exact inference from a current assignment and the remaining schematic. We show how our bounding method can be used within a lifted importance sampling algorithm, in order to perform effective Rao-Blackwellisation, and demonstrate experimentally that the Rao-Blackwellised version of the algorithm yields more accurate estimates on several real-world datasets.

****

Mixed Robust/Average Submodular Partitioning: Fast Algorithms, Guarantees, and Applications

Kai Wei, Rishabh K. Iyer, Shengjie Wang, Wenruo Bai, Jeff A. Bilmes

We investigate two novel mixed robust/average-case submodular data partitioning problems that we collectively call Submodular Partitioning. These problems generalize purely robust instances of the problem, namely max-min submodular fair allocation (SFA) and \emph{min-max submodular load balancing} (SLB), and also average-case instances, that is the submodular welfare problem (SWP) and submodular multiway partition (SMP). While the robust versions have been studied in the theory community, existing work has focused on tight approximation guarantees, and the resultant algorithms are not generally scalable to large real-world applications. This contrasts the average case instances, where most of the algorithms are scalable. In the present paper, we bridge this gap, by proposing several new algorithms (including greedy, majorization-minimization, minorization-maximization, and relaxation algorithms) that not only scale to large datasets but that also achieve theoretical approximation guarantees comparable to the state-of-the-art. We moreover provide new scalable algorithms that apply to additive combinations of the robust and average-case objectives. We show that these problems have m any applications in machine learning (ML), including data partitioning and load balancing for distributed ML, data clustering, and image segmentation. We empirically demonstrate the efficacy of our algorithms on real-world problems involving data partitioning for distributed optimization (of convex and deep neural network objectives), and also purely unsupervised image segmentation.

****

Gradient Estimation Using Stochastic Computation Graphs

John Schulman, Nicolas Heess, Theophane Weber, Pieter Abbeel

In a variety of problems originating in supervised, unsupervised, and reinforcement learning, the loss function is defined by an expectation over a collection of random variables, which might be part of a probabilistic model or the external world. Estimating the gradient of this loss function, using samples, lies at the core of gradient-based learning algorithms for these problems. We introduce the formalism of stochastic computation graphs--directed acyclic graphs that include both deterministic functions and conditional probability distributions and describe how to easily and automatically derive an unbiased estimator of the loss function's gradient. The resulting algorithm for computing the gradient estimator is a simple modification of the standard backpropagation algorithm. The generic scheme we propose unifies estimators derived in variety of prior work, along with variance-reduction techniques therein. It could assist researchers in developing intricate models involving a combination of stochastic and deterministic operations, enabling, for example, attention, memory, and control actions.

****

Rectified Factor Networks

Djork-Arné Clevert, Andreas Mayr, Thomas Unterthiner, Sepp Hochreiter

We propose rectified factor networks (RFNs) to efficiently construct very sparse, non-linear, high-dimensional representations of the input. RFN models identify rare and small events, have a low interference between code units, have a small reconstruction error, and explain the data covariance structure. RFN learning is a generalized alternating minimization algorithm derived from the posterior regularization method which enforces non-negative and normalized posterior means. We proof convergence and correctness of the RFN learning algorithm.On benchmarks, RFNs are compared to other unsupervised methods like autoencoders, RBMs, factor analysis, ICA, and PCA. In contrast to previous sparse coding methods, RFNs yield sparser codes, capture the data's covariance structure more precisely, and have a significantly smaller reconstruction error. We test RFNs as pretraining technique of deep networks on different vision datasets, where RFNs were superior to RBMs and autoencoders. On gene expression data from two pharmaceutical drug discovery studies, RFNs detected small and rare gene modules that revealed highly relevant new biological insights which were so far missed by other unsupervised methods.RFN package for GPU/CPU is available at http://www.bioinf.jku.at/software/rfn.

****

Adaptive Stochastic Optimization: From Sets to Paths

Zhan Wei Lim, David Hsu, Wee Sun Lee

Adaptive stochastic optimization optimizes an objective function adaptively under uncertainty. Adaptive stochastic optimization plays a crucial role in planning and learning under uncertainty, but is, unfortunately, computationally intractable in general. This paper introduces two conditions on the objective function, the marginal likelihood rate bound and the marginal likelihood bound, which enable efficient approximate solution of adaptive stochastic optimization. Several interesting classes of functions satisfy these conditions naturally, e.g., the version space reduction function for hypothesis learning. We describe Recursive Adaptive Coverage (RAC), a new adaptive stochastic optimization algorithm that exploits these conditions, and apply it to two planning tasks under uncertainty. In constrast to the earlier submodular optimization approach, our algorithm applies to adaptive stochastic optimization algorithm over both sets and paths.

****

A Universal Primal-Dual Convex Optimization Framework

Alp Yurtsever, Quoc Tran Dinh, Volkan Cevher

We propose a new primal-dual algorithmic framework for a prototypical constrained convex optimization template. The algorithmic instances of our framework are universal since they can automatically adapt to the unknown Holder continuity degree and constant within the dual formulation. They are also guaranteed to have optimal convergence rates in the objective residual and the feasibility gap for each Holder smoothness degree. In contrast to existing primal-dual algorithms, our framework avoids the proximity operator of the objective function. We instead leverage computationally cheaper, Fenchel-type operators, which are the main workhorses of the generalized conditional gradient (GCG)-type methods. In contrast to the GCG-type methods, our framework does not require the objective function to be differentiable, and can also process additional general linear inclusion constraints, while guarantees the convergence rate on the primal problem.

****

Adversarial Prediction Games for Multivariate Losses
Hong Wang, Wei Xing, Kaiser Asif, Brian Ziebart

Multivariate loss functions are used to assess performance in many modern prediction tasks, including information retrieval and ranking applications. Convex approximations are typically optimized in their place to avoid NP-hard empirical risk minimization problems. We propose to approximate the training data instead of the loss function by posing multivariate prediction as an adversarial game between a loss-minimizing prediction player and a loss-maximizing evaluation player constrained to match specified properties of training data. This avoids the non-convexity of empirical risk minimization, but game sizes are exponential in the number of predicted variables. We overcome this intractability using the double oracle constraint generation method. We demonstrate the efficiency and predictive performance of our approach on tasks evaluated using the precision at k, the F-score and the discounted cumulative gain.

****

Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning
Shakir Mohamed, Danilo Jimenez Rezende

The mutual information is a core statistical quantity that has applications in all areas of machine learning, whether this is in training of density models over multiple data modalities, in maximising the efficiency of noisy transmission channels, or when learning behaviour policies for exploration by artificial agents. Most learning algorithms that involve optimisation of the mutual information rely on the Blahut-Arimoto algorithm --- an enumerative algorithm with exponential complexity that is not suitable for modern machine learning applications. This paper provides a new approach for scalable optimisation of the mutual information by merging techniques from variational inference and deep learning. We develop our approach by focusing on the problem of intrinsically-motivated learning, where the mutual information forms the definition of a well-known internal drive known as empowerment. Using a variational lower bound on the mutual information, combined with convolutional networks for handling visual input streams, we develop a stochastic optimisation algorithm that allows for scalable information maximisation and empowerment-based reasoning directly from pixels to actions.

****

Deep Visual Analogy-Making
Scott E. Reed, Yi Zhang, Yuting Zhang, Honglak Lee

In addition to identifying the content within a single image, relating images and generating related images are critical tasks for image understanding. Recently, deep convolutional networks have yielded breakthroughs in producing image labels, annotations and captions, but have only just begun to be used for producing high-quality image outputs. In this paper we develop a novel deep network trained end-to-end to perform visual analogy making, which is the task of transforming a query image according to an example pair of related images. Solving this problem requires both accurately recognizing a visual relationship and generating a transformed query image accordingly. Inspired by recent advances in language modeling, we propose to solve visual analogies by learning to map images to a neural embedding in which analogical reasoning is simple, such as by vector subtracti on and addition. In experiments, our model effectively models visual analogies on several datasets: 2D shapes, animated video game sprites, and 3D car models.

****

Rate-Agnostic (Causal) Structure Learning

Sergey Plis, David Danks, Cynthia Freeman, Vince Calhoun

Causal structure learning from time series data is a major scientific challenge. Existing algorithms assume that measurements occur sufficiently quickly; more precisely, they assume that the system and measurement timescales are approximately equal. In many scientific domains, however, measurements occur at a significantly slower rate than the underlying system changes. Moreover, the size of the mismatch between timescales is often unknown. This paper provides three distinct causal structure learning algorithms, all of which discover all dynamic graphs that could explain the observed measurement data as arising from undersampling at some rate. That is, these algorithms all learn causal structure without assuming any particular relation between the measurement and system timescales; they are thus rate-agnostic. We apply these algorithms to data from simulations. The results provide insight into the challenge of undersampling.

****

Structured Estimation with Atomic Norms: General Bounds and Applications

Sheng Chen, Arindam Banerjee

For structured estimation problems with atomic norms, recent advances in the literature express sample complexity and estimation error bounds in terms of certain geometric measures, in particular Gaussian width of the unit norm ball, Gaussian width of a spherical cap induced by a tangent cone, and a restricted norm compatibility constant. However, given an atomic norm, bounding these geometric measures can be difficult. In this paper, we present general upper bounds for such geometric measures, which only require simple information of the atomic norm under consideration, and we establish tightness of these bounds by providing the corresponding lower bounds. We show applications of our analysis to certain atomic norms, especially k-support norm, for which existing result is incomplete.

****

Logarithmic Time Online Multiclass prediction

Anna E. Choromanska, John Langford

We study the problem of multiclass classification with an extremely large number of classes (k), with the goal of obtaining train and test time complexity logarithmic in the number of classes. We develop top-down tree construction approaches for constructing logarithmic depth trees. On the theoretical front, we formulate a new objective function, which is optimized at each node of the tree and creates dynamic partitions of the data which are both pure (in terms of class labels) and balanced. We demonstrate that under favorable conditions, we can construct logarithmic depth trees that have leaves with low label entropy. However, the objective function at the nodes is challenging to optimize computationally. We address the empirical problem with a new online decision tree construction procedure. Experiments demonstrate that this online algorithm quickly achieves improvement in test error compared to more common logarithmic training time approaches, which makes it a plausible method in computationally constrained large-k applications.

****

Copula variational inference

Dustin Tran, David Blei, Edo M. Airoldi

We develop a general variational inference method that preserves dependency among the latent variables. Our method uses copulas to augment the families of distributions used in mean-field and structured approximations. Copulas model the dependency that is not captured by the original variational distribution, and thus the augmented variational family guarantees better approximations to the posterior. With stochastic optimization, inference on the augmented distribution is scalable. Furthermore, our strategy is generic: it can be applied to any inference procedure that currently uses the mean-field or structured approach. Copula variational inference has many advantages: it reduces bias; it is less sensitive to local optima; it is less sensitive to hyperparameters; and it helps characterize and interpret the dependency among the latent variables.

****

Attractor Network Dynamics Enable Preplay and Rapid Path Planning in Maze–like Environments

Dane S. Corneil, Wulfram Gerstner

Rodents navigating in a well-known environment can rapidly learn and revisit observed reward locations, often after a single trial. While the mechanism for rapid path planning is unknown, the CA3 region in the hippocampus plays an important role, and emerging evidence suggests that place cell activity during hippocampal preplay periods may trace out future goal-directed trajectories. Here, we show how a particular mapping of space allows for the immediate generation of trajectories between arbitrary start and goal locations in an environment, based only on the mapped representation of the goal. We show that this representation can be implemented in a neural attractor network model, resulting in bump--like activity profiles resembling those of the CA3 region of hippocampus. Neurons tend to locally excite neurons with similar place field centers, while inhibiting other neurons with distant place field centers, such that stable bumps of activity can form at arbitrary locations in the environment. The network is initialized to represent a point in the environment, then weakly stimulated with an input corresponding to an arbitrary goal location. We show that the resulting activity can be interpreted as a gradient ascent on the value function induced by a reward at the goal location. Indeed, in networks with large place fields, we show that the network properties cause the bump to move smoothly from its initial location to the goal, around obstacles or walls. Our results illustrate that an attractor network with hippocampal-like attributes may be important for rapid path planning.

****

Exactness of Approximate MAP Inference in Continuous MRFs

Nicholas Ruozzi

Computing the MAP assignment in graphical models is generally intractable. As a result, for discrete graphical models, the MAP problem is often approximated using linear programming relaxations. Much research has focused on characterizing when these LP relaxations are tight, and while they are relatively well-understood in the discrete case, only a few results are known for their continuous analog. In this work, we use graph covers to provide necessary and sufficient conditions for continuous MAP relaxations to be tight. We use this characterization to give simple proofs that the relaxation is tight for log-concave decomposable and log-supermodular decomposable models. We conclude by exploring the relationship between these two seemingly distinct classes of functions and providing specific conditions under which the MAP relaxation can and cannot be tight.

****

Explore no more: Improved high-probability regret bounds for non-stochastic bandits

Gergely Neu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Subspace Clustering with Irrelevant Features via Robust Dantzig Selector

Chao Qu, Huan Xu

This paper considers the subspace clustering problem where the data contains irrelevant or corrupted features. We propose a method termed ``robust Dantzig selector'' which can successfully identify the clustering structure even with the presence of irrelevant features. The idea is simple yet powerful: we replace the inner product by its robust counterpart, which is insensitive to the irrelevant features given an upper bound of the number of irrelevant features. We establish theoretical guarantees for the algorithm to identify the correct subspace, and demonstrate the effectiveness of the algorithm via numerical simulations. To the best of our knowledge, this is the first method developed to tackle subspace clus tering with irrelevant features.

****

Variational Consensus Monte Carlo
Maxim Rabinovich, Elaine Angelino, Michael I. Jordan

Practitioners of Bayesian statistics have long depended on Markov chain Monte Carlo (MCMC) to obtain samples from intractable posterior distributions. Unfortunately, MCMC algorithms are typically serial, and do not scale to the large datasets typical of modern machine learning. The recently proposed consensus Monte Carlo algorithm removes this limitation by partitioning the data and drawing samples conditional on each partition in parallel (Scott et al, 2013). A fixed aggregation function then combines these samples, yielding approximate posterior samples. We introduce variational consensus Monte Carlo (VCMC), a variational Bayes algorithm that optimizes over aggregation functions to obtain samples from a distribution that better approximates the target. The resulting objective contains an intractable entropy term; we therefore derive a relaxation of the objective and show that the relaxed problem is blockwise concave under mild conditions. We illustrate the advantages of our algorithm on three inference tasks from the literature, demonstrating both the superior quality of the posterior approximation and the moderate overhead of the optimization step. Our algorithm achieves a relative error reduction (measured against serial MCMC) of up to 39% compared to consensus Monte Carlo on the task of estimating 300-dimensional probit regression parameter expectations; similarly, it achieves an error reduction of 92% on the task of estimating cluster comembership probabilities in a Gaussian mixture model with 8 components in 8 dimensions. Furthermore, these gains come at moderate cost compared to the runtime of serial MCMC, achieving near-ideal speedup in some instances.

****

Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks
Samy Bengio, Oriol Vinyals, Navdeep Jaitly, Noam Shazeer

Recurrent Neural Networks can be trained to produce sequences of tokens given some input, as exemplified by recent results in machine translation and image captioning. The current approach to training them consists of maximizing the likelihood of each token in the sequence given the current (recurrent) state and the previous token. At inference, the unknown previous token is then replaced by a token generated by the model itself. This discrepancy between training and inference can yield errors that can accumulate quickly along the generated sequence. We propose a curriculum learning strategy to gently change the training process from a fully guided scheme using the true previous token, towards a less guided scheme which mostly uses the generated token instead. Experiments on several sequence prediction tasks show that this approach yields significant improvements. Moreover, it was used successfully in our winning bid to the MSCOCO image captioning challenge, 2015.

****

Path-SGD: Path-Normalized Optimization in Deep Neural Networks
Behnam Neyshabur, Russ R. Salakhutdinov, Nati Srebro

We revisit the choice of SGD for training deep neural networks by reconsidering the appropriate geometry in which to optimize the weights. We argue for a geometry invariant to rescaling of weights that does not affect the output of the network, and suggest Path-SGD, which is an approximate steepest descent method with respect to a path-wise regularizer related to max-norm regularization. Path-SGD is easy and efficient to implement and leads to empirical gains over SGD and AdaGrad.

****

Orthogonal NMF through Subspace Exploration
Megasthenis Asteris, Dimitris Papailiopoulos, Alexandros G. Dimakis

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

M-Statistic for Kernel Change-Point Detection

Shuang Li, Yao Xie, Hanjun Dai, Le Song

Detecting the emergence of an abrupt change-point is a classic problem in statistics and machine learning. Kernel-based nonparametric statistics have been proposed for this task which make fewer assumptions on the distributions than traditional parametric approach. However, none of the existing kernel statistics has provided a computationally efficient way to characterize the extremal behavior of the statistic. Such characterization is crucial for setting the detection threshold, to control the significance level in the offline case as well as the average run length in the online case. In this paper we propose two related computationally efficient M-statistics for kernel-based change-point detection when the amount of background data is large. A novel theoretical result of the paper is the characterization of the tail probability of these statistics using a new technique based on change-of-measure. Such characterization provides us accurate detection thresholds for both offline and online cases in computationally efficient manner, without the need to resort to the more expensive simulations such as bootstrapping. We show that our methods perform well in both synthetic and real world data.

****

Active Learning from Weak and Strong Labelers

Chicheng Zhang, Kamalika Chaudhuri

An active learner is given a hypothesis class, a large set of unlabeled examples and the ability to interactively query labels to an oracle of a subset of these examples; the goal of the learner is to learn a hypothesis in the class that fits the data well by making as few label queries as possible.This work addresses active learning with labels obtained from strong and weak labelers, where in addition to the standard active learning setting, we have an extra weak labeler which may occasionally provide incorrect labels. An example is learning to classify medical images where either expensive labels may be obtained from a physician (oracle or strong labeler), or cheaper but occasionally incorrect labels may be obtained from a medical resident (weak labeler). Our goal is to learn a classifier with low error on data labeled by the oracle, while using the weak labeler to reduce the number of label queries made to this labeler. We provide an active learning algorithm for this setting, establish its statistical consistency, and analyze its label complexity to characterize when it can provide label savings over using the strong labeler alone.

****

Sum-of-Squares Lower Bounds for Sparse PCA

Tengyu Ma, Avi Wigderson

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Where are they looking?

Adria Recasens, Aditya Khosla, Carl Vondrick, Antonio Torralba

Humans have the remarkable ability to follow the gaze of other people to identify what they are looking at. Following eye gaze, or gaze-following, is an important ability that allows us to understand what other people are thinking, the actions they are performing, and even predict what they might do next. Despite the importance of this topic, this problem has only been studied in limited scenarios within the computer vision community. In this paper, we propose a deep neural network-based approach for gaze-following and a new benchmark dataset for thorough evaluation. Given an image and the location of a head, our approach follows the gaze of the person and identifies the object being looked at. After training, the network is able to discover how to extract head pose and gaze orientation, and to select objects in the scene that are in the predicted line of sight and likely to be looked at (such as televisions, balls and food). The quantitative evaluation shows that our approach produces reliable results, even when viewing only the back of the head. While our method outperforms several baseline approache s, we are still far from reaching human performance at this task. Overall, we believe that this is a challenging and important task that deserves more attention from the community.

****

Softstar: Heuristic-Guided Probabilistic Inference

Mathew Monfort, Brenden M. Lake, Brenden M. Lake, Brian Ziebart, Patrick Lucey, Josh Tenenbaum

Recent machine learning methods for sequential behavior prediction estimate the motives of behavior rather than the behavior itself. This higher-level abstraction improves generalization in different prediction settings, but computing predictions often becomes intractable in large decision spaces. We propose the Softstar algorithm, a softened heuristic-guided search technique for the maximum entropy inverse optimal control model of sequential behavior. This approach supports probabilistic search with bounded approximation error at a significantly reduced computational cost when compared to sampling based methods. We present the algorithm, analyze approximation guarantees, and compare performance with simulation-based inference on two distinct complex decision tasks.

****

Fast Bidirectional Probability Estimation in Markov Models

Siddhartha Banerjee, Peter Lofgren

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Learning to Linearize Under Uncertainty

Ross Goroshin, Michael F. Mathieu, Yann LeCun

Training deep feature hierarchies to solve supervised learning tasks has achieving state of the art performance on many problems in computer vision. However, a principled way in which to train such hierarchies in the unsupervised setting has remained elusive. In this work we suggest a new architecture and loss for training deep feature hierarchies that linearize the transformations observed in unlabelednatural video sequences. This is done by training a generative model to predict video frames. We also address the problem of inherent uncertainty in prediction by introducing a latent variables that are non-deterministic functions of the input into the network architecture.

****

Variance Reduced Stochastic Gradient Descent with Neighbors

Thomas Hofmann, Aurelien Lucchi, Simon Lacoste-Julien, Brian McWilliams

Stochastic Gradient Descent (SGD) is a workhorse in machine learning, yet it is also known to be slow relative to steepest descent. Recently, variance reduction techniques such as SVRG and SAGA have been proposed to overcome this weakness. With asymptotically vanishing variance, a constant step size can be maintained, resulting in geometric convergence rates. However, these methods are either based on occasional computations of full gradients at pivot points (SVRG), or on keeping per data point corrections in memory (SAGA). This has the disadvantage that one cannot employ these methods in a streaming setting and that speed-ups relative to SGD may need a certain number of epochs in order to materialize. This paper investigates a new class of algorithms that can exploit neighborhood structure in the training data to share and re-use information about past stochastic gradients across data points. While not meant to be offering advantages in an asymptotic setting, there are significant benefits in the transient optimization phase, in particular in a streaming or single-epoch setting. We investigate this family of algorithms in a thorough analysis and show supporting experimental results. As a side-product we provide a simple and unified proof technique for a broad class of variance reduction algorithms.

****

On Elicitation Complexity

Rafael Frongillo, Ian Kash

Elicitation is the study of statistics or properties which are computable via em pirical risk minimization. While several recent papers have approached the general question of which properties are elicitable, we suggest that this is the wrong question---all properties are elicitable by first eliciting the entire distribution or data set, and thus the important question is how elicitable. Specifically, what is the minimum number of regression parameters needed to compute the property?Building on previous work, we introduce a new notion of elicitation complexity and lay the foundations for a calculus of elicitation. We establish several general results and techniques for proving upper and lower bounds on elicitation complexity. These results provide tight bounds for eliciting the Bayes risk of any loss, a large class of properties which includes spectral risk measures and several new properties of interest.

****

Learning with Relaxed Supervision

Jacob Steinhardt, Percy S. Liang

For weakly-supervised problems with deterministic constraints between the latent variables and observed output, learning necessitates performing inference over latent variables conditioned on the output, which can be intractable no matter how simple the model family is. Even finding a single latent variable setting that satisfies the constraints could be difficult; for instance, the observed output may be the result of a latent database query or graphics program which must be inferred. Here, the difficulty lies in not the model but the supervision, and poor approximations at this stage could lead to following the wrong learning signal entirely. In this paper, we develop a rigorous approach to relaxing the supervision, which yields asymptotically consistent parameter estimates despite altering the supervision. Our approach parameterizes a family of increasingly accurate relaxations, and jointly optimizes both the model and relaxation parameters, while formulating constraints between these parameters to ensure efficient inference. These efficiency constraints allow us to learn in otherwise intractable settings, while asymptotic consistency ensures that we always follow a valid learning signal.

****

Matrix Completion Under Monotonic Single Index Models

Ravi Sastry Ganti, Laura Balzano, Rebecca Willett

Most recent results in matrix completion assume that the matrix under consideration is low-rank or that the columns are in a union of low-rank subspaces. In real-world settings, however, the linear structure underlying these models is distorted by a (typically unknown) nonlinear transformation. This paper addresses the challenge of matrix completion in the face of such nonlinearities. Given a few observations of a matrix that are obtained by applying a Lipschitz, monotonic function to a low rank matrix, our task is to estimate the remaining unobserved entries. We propose a novel matrix completion method that alternates between low-rank matrix estimation and monotonic function estimation to estimate the missing matrix elements. Mean squared error bounds provide insight into how well the matrix can be estimated based on the size, rank of the matrix and properties of the nonlinear transformation. Empirical results on synthetic and real-world datasets demonstrate the competitiveness of the proposed approach.

****

Principal Geodesic Analysis for Probability Measures under the Optimal Transport Metric

Vivien Seguy, Marco Cuturi

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

HONOR: Hybrid Optimization for NOn-convex Regularized problems

Pinghua Gong, Jieping Ye

Recent years have witnessed the superiority of non-convex sparse learning formulations over their convex counterparts in both theory and practice. However, due to the non-convexity and non-smoothness of the regularizer, how to efficiently s

olve the non-convex optimization problem for large-scale data is still quite challenging. In this paper, we propose an efficient \underline{H}ybrid \underline{O}ptimization algorithm for \underline{NO}n convex \underline{R}egularized problems (HONOR). Specifically, we develop a hybrid scheme which effectively integrates a Quasi-Newton (QN) step and a Gradient Descent (GD) step. Our contributions are as follows: (1) HONOR incorporates the second-order information to greatly speed up the convergence, while it avoids solving a regularized quadratic programming and only involves matrix-vector multiplications without explicitly forming the inverse Hessian matrix. (2) We establish a rigorous convergence analysis for HONOR, which shows that convergence is guaranteed even for non-convex problems, while it is typically challenging to analyze the convergence for non-convex problems. (3) We conduct empirical studies on large-scale data sets and results demonstrate that HONOR converges significantly faster than state-of-the-art algorithms.
****
The Poisson Gamma Belief Network
Mingyuan Zhou, Yulai Cong, Bo Chen
To infer a multilayer representation of high-dimensional count vectors, we propose the Poisson gamma belief network (PGBN) that factorizes each of its layers into the product of a connection weight matrix and the nonnegative real hidden units of the next layer. The PGBN's hidden layers are jointly trained with an upward-downward Gibbs sampler, each iteration of which upward samples Dirichlet distributed connection weight vectors starting from the first layer (bottom data layer), and then downward samples gamma distributed hidden units starting from the top hidden layer. The gamma-negative binomial process combined with a layer-wise training strategy allows the PGBN to infer the width of each layer given a fixed budget on the width of the first layer. The PGBN with a single hidden layer reduces to Poisson factor analysis. Example results on text analysis illustrate interesting relationships between the width of the first layer and the inferred network structure, and demonstrate that the PGBN, whose hidden units are imposed with correlated gamma priors, can add more layers to increase its performance gains over Poisson factor analysis, given the same limit on the width of the first layer.
****
Fixed-Length Poisson MRF: Adding Dependencies to the Multinomial
David I. Inouye, Pradeep K. Ravikumar, Inderjit S. Dhillon
We propose a novel distribution that generalizes the Multinomial distribution to enable dependencies between dimensions. Our novel distribution is based on the parametric form of the Poisson MRF model [Yang et al., 2012] but is fundamentally different because of the domain restriction to a fixed-length vector like in a Multinomial where the number of trials is fixed or known. Thus, we propose the Fixed-Length Poisson MRF (LPMRF) distribution. We develop methods to estimate the likelihood and log partition function (i.e. the log normalizing constant), which was not developed for the Poisson MRF model. In addition, we propose novel mixture and topic models that use LPMRF as a base distribution and discuss the similarities and differences with previous topic models such as the recently proposed Admixture of Poisson MRFs [Inouye et al., 2014]. We show the effectiveness of our LPMRF distribution over Multinomial models by evaluating the test set perplexity on a dataset of abstracts and Wikipedia. Qualitatively, we show that the positive dependencies discovered by LPMRF are interesting and intuitive. Finally, we show that our algorithms are fast and have good scaling (code available online).
****
Stochastic Expectation Propagation
Yingzhen Li, José Miguel Hernández-Lobato, Richard E. Turner
Expectation propagation (EP) is a deterministic approximation algorithm that is often used to perform approximate Bayesian parameter learning. EP approximates the full intractable posterior distribution through a set of local-approximations that are iteratively refined for each datapoint. EP can offer analytic and computational advantages over other approximations, such as Variational Inference (V

I), and is the method of choice for a number of models. The local nature of EP appears to make it an ideal candidate for performing Bayesian learning on large models in large-scale datasets settings. However, EP has a crucial limitation in this context: the number approximating factors needs to increase with the number of data-points, N, which often entails a prohibitively large memory overhead. This paper presents an extension to EP, called stochastic expectation propagation (SEP), that maintains a global posterior approximation (like VI) but updates it in a local way (like EP). Experiments on a number of canonical learning problems using synthetic and real-world datasets indicate that SEP performs almost as well as full EP, but reduces the memory consumption by a factor of N. SEP is therefore ideally suited to performing approximate Bayesian learning in the large model, large dataset setting.

****

Beyond Sub-Gaussian Measurements: High-Dimensional Structured Estimation with Sub-Exponential Designs

Vidyashankar Sivakumar, Arindam Banerjee, Pradeep K. Ravikumar

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Fast Randomized Kernel Ridge Regression with Statistical Guarantees

Ahmed Alaoui, Michael W. Mahoney

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Skip-Thought Vectors

Ryan Kiros, Yukun Zhu, Russ R. Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, Sanja Fidler

We describe an approach for unsupervised learning of a generic, distributed sentence encoder. Using the continuity of text from books, we train an encoder-decoder model that tries to reconstruct the surrounding sentences of an encoded passage. Sentences that share semantic and syntactic properties are thus mapped to similar vector representations. We next introduce a simple vocabulary expansion method to encode words that were not seen as part of training, allowing us to expand our vocabulary to a million words. After training our model, we extract and evaluate our vectors with linear models on 8 tasks: semantic relatedness, paraphrase detection, image-sentence ranking, question-type classification and 4 benchmark sentiment and subjectivity datasets. The end result is an off-the-shelf encoder that can produce highly generic sentence representations that are robust and perform well in practice. We will make our encoder publicly available.

****

Collaborative Filtering with Graph Information: Consistency and Scalable Methods

Nikhil Rao, Hsiang-Fu Yu, Pradeep K. Ravikumar, Inderjit S. Dhillon

Low rank matrix completion plays a fundamental role in collaborative filtering applications, the key idea being that the variables lie in a smaller subspace than the ambient space. Often, additional information about the variables is known, and it is reasonable to assume that incorporating this information will lead to better predictions. We tackle the problem of matrix completion when pairwise relationships among variables are known, via a graph. We formulate and derive a highly efficient, conjugate gradient based alternating minimization scheme that solves optimizations with over 55 million observations up to 2 orders of magnitude faster than state-of-the-art (stochastic) gradient-descent based methods. On the theoretical front, we show that such methods generalize weighted nuclear norm formulations, and derive statistical consistency guarantees. We validate our results on both real and synthetic datasets.

****

Gaussian Process Random Fields

David Moore, Stuart J. Russell

Gaussian processes have been successful in both supervised and unsupervised machine learning tasks, but their computational complexity has constrained practical applications. We introduce a new approximation for large-scale Gaussian processes, the Gaussian Process Random Field (GPRF), in which local GPs are coupled via pairwise potentials. The GPRF likelihood is a simple, tractable, and parallelizeable approximation to the full GP marginal likelihood, enabling latent variable modeling and hyperparameter selection on large datasets. We demonstrate its effectiveness on synthetic spatial data as well as a real-world application to seismic event location.

****

Decoupled Deep Neural Network for Semi-supervised Semantic Segmentation

Seunghoon Hong, Hyeonwoo Noh, Bohyung Han

We propose a novel deep neural network architecture for semi-supervised semantic segmentation using heterogeneous annotations. Contrary to existing approaches posing semantic segmentation as region-based classification, our algorithm decouples classification and segmentation, and learns a separate network for each task. In this architecture, labels associated with an image are identified by classification network, and binary segmentation is subsequently performed for each identified label by segmentation network. The decoupled architecture enables us to learn classification and segmentation networks separately based on the training data with image-level and pixel-wise class labels, respectively. It facilitates to reduce search space for segmentation effectively by exploiting class-specific activation maps obtained from bridging layers. Our algorithm shows outstanding performance compared to other semi-supervised approaches even with much less training images with strong annotations in PASCAL VOC dataset.

****

Discrete Rényi Classifiers

Meisam Razaviyayn, Farzan Farnia, David Tse

Consider the binary classification problem of predicting a target variable Y from a discrete feature vector X = (X1,...,Xd). When the probability distribution P(X,Y) is known, the optimal classifier, leading to the minimum misclassification rate, is given by the Maximum A-posteriori Probability (MAP) decision rule. However, in practice, estimating the complete joint distribution P(X,Y) is computationally and statistically impossible for large values of d. Therefore, an alternative approach is to first estimate some low order marginals of the joint probability distribution P(X,Y) and then design the classifier based on the estimated low order marginals. This approach is also helpful when the complete training data instances are not available due to privacy concerns. In this work, we consider the problem of designing the optimum classifier based on some estimated low order marginals of (X,Y). We prove that for a given set of marginals, the minimum Hirschfeld-Gebelein-R´enyi (HGR) correlation principle introduced in [1] leads to a randomized classification rule which is shown to have a misclassification rate no larger than twice the misclassification rate of the optimal classifier. Then, we show that under a separability condition, the proposed algorithm is equivalent to a randomized linear regression approach which naturally results in a robust feature selection method selecting a subset of features having the maximum worst case HGR correlation with the target variable. Our theoretical upper-bound is similar to the recent Discrete Chebyshev Classifier (DCC) approach [2], while the proposed algorithm has significant computational advantages since it only requires solving a least square optimization problem. Finally, we numerically compare our proposed algorithm with the DCC classifier and show that the proposed algorithm results in better misclassification rate over various UCI data repository datasets.

****

Preconditioned Spectral Descent for Deep Learning

David E. Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, Volkan Cevher

Deep learning presents notorious computational challenges. These challenges include, but are not limited to, the non-convexity of learning objectives and estimating the quantities needed for optimization algorithms, such as gradients. While we do not address the non-convexity, we present an optimization solution that ex- ploits the so far unused “geometry” in the objective function in order to best make use of the estimated gradients. Previous work attempted similar goals with preconditioned methods in the Euclidean space, such as L-BFGS, RMSprop, and ADA-grad. In stark contrast, our approach combines a non-Euclidean gradient method with preconditioning. We provide evidence that this combination more accurately captures the geometry of the objective function compared to prior work. We theoretically formalize our arguments and derive novel preconditioned non-Euclidean algorithms. The results are promising in both computational time and quality when applied to Restricted Boltzmann Machines, Feedforward Neural Nets, and Convolutional Neural Nets.

****

Accelerated Mirror Descent in Continuous and Discrete Time
Walid Krichene, Alexandre Bayen, Peter L. Bartlett

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Accelerated Proximal Gradient Methods for Nonconvex Programming
Huan Li, Zhouchen Lin

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Monotone k-Submodular Function Maximization with Size Constraints
Naoto Ohsaka, Yuichi Yoshida

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

****

Spherical Random Features for Polynomial Kernels
Jeffrey Pennington, Felix Xinnan X. Yu, Sanjiv Kumar

Compact explicit feature maps provide a practical framework to scale kernel methods to large-scale learning, but deriving such maps for many types of kernels remains a challenging open problem. Among the commonly used kernels for nonlinear classification are polynomial kernels, for which low approximation error has thus far necessitated explicit feature maps of large dimensionality, especially for higher-order polynomials. Meanwhile, because polynomial kernels are unbounded, they are frequently applied to data that has been normalized to unit l2 norm. The question we address in this work is: if we know a priori that data is so normalized, can we devise a more compact map? We show that a putative affirmative answer to this question based on Random Fourier Features is impossible in this setting, and introduce a new approximation paradigm, Spherical Random Fourier (SRF) features, which circumvents these issues and delivers a compact approximation to polynomial kernels for data on the unit sphere. Compared to prior work, SRF features are less rank-deficient, more compact, and achieve better kernel approximation, especially for higher-order polynomials. The resulting predictions have lower variance and typically yield better classification accuracy.

****

A Dual Augmented Block Minimization Framework for Learning with Limited Memory
Ian En-Hsu Yen, Shan-Wei Lin, Shou-De Lin

In past few years, several techniques have been proposed for training of linear Support Vector Machine (SVM) in limited-memory setting, where a dual block-coordinate descent (dual-BCD) method was used to balance cost spent on I/O and computation. In this paper, we consider the more general setting of regularized \emph{Empirical Risk Minimization (ERM)} when data cannot fit into memory. In particular, we generalize the existing block minimization framework based on strong dual ity and \emph{Augmented Lagrangian} technique to achieve global convergence for ERM with arbitrary convex loss function and regularizer. The block minimization framework is flexible in the sense that, given a solver working under sufficient memory, one can integrate it with the framework to obtain a solver globally convergent under limited-memory condition. We conduct experiments on L1-regularized classification and regression problems to corroborate our convergence theory and compare the proposed framework to algorithms adopted from online and distributed settings, which shows superiority of the proposed approach on data of size ten times larger than the memory capacity.

****

Convolutional Networks on Graphs for Learning Molecular Fingerprints

David K. Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan Aspuru-Guzik, Ryan P. Adams

We introduce a convolutional neural network that operates directly on graphs.These networks allow end-to-end learning of prediction pipelines whose inputs are graphs of arbitrary size and shape.The architecture we present generalizes standard molecular feature extraction methods based on circular fingerprints.We show that these data-driven features are more interpretable, and have better predictive performance on a variety of tasks.

****

Decomposition Bounds for Marginal MAP

Wei Ping, Qiang Liu, Alexander T. Ihler

Marginal MAP inference involves making MAP predictions in systems defined with latent variables or missing information. It is significantly more difficult than pure marginalization and MAP tasks, for which a large class of efficient and convergent variational algorithms, such as dual decomposition, exist. In this work, we generalize dual decomposition to a generic powered-sum inference task, which includes marginal MAP, along with pure marginalization and MAP, as special cases. Our method is based on a block coordinate descent algorithm on a new convex decomposition bound, that is guaranteed to converge monotonically, and can be parallelized efficiently. We demonstrate our approach on various inference queries over real-world problems from the UAI approximate inference challenge, showing that our framework is faster and more reliable than previous methods.

****

The Brain Uses Reliability of Stimulus Information when Making Perceptual Decisions

Sebastian Bitzer, Stefan Kiebel

In simple perceptual decisions the brain has to identify a stimulus based on noisy sensory samples from the stimulus. Basic statistical considerations state that the reliability of the stimulus information, i.e., the amount of noise in the samples, should be taken into account when the decision is made. However, for perceptual decision making experiments it has been questioned whether the brain indeed uses the reliability for making decisions when confronted with unpredictable changes in stimulus reliability. We here show that even the basic drift diffusion model, which has frequently been used to explain experimental findings in perceptual decision making, implicitly relies on estimates of stimulus reliability. We then show that only those variants of the drift diffusion model which allow stimulus-specific reliabilities are consistent with neurophysiological findings. Our analysis suggests that the brain estimates the reliability of the stimulus on a short time scale of at most a few hundred milliseconds.

****

Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, Wei Xu

In this paper, we present the mQA model, which is able to answer questions about the content of an image. The answer can be a sentence, a phrase or a single word. Our model contains four components: a Long Short-Term Memory (LSTM) to extract the question representation, a Convolutional Neural Network (CNN) to extract the visual representation, an LSTM for storing the linguistic context in an answer, and a fusing component to combine the information from the first three compon ents and generate the answer. We construct a Freestyle Multilingual Image Question Answering (FM-IQA) dataset to train and evaluate our mQA model. It contains over 150,000 images and 310,000 freestyle Chinese question-answer pairs and their English translations. The quality of the generated answers of our mQA model on this dataset is evaluated by human judges through a Turing Test. Specifically, we mix the answers provided by humans and our model. The human judges need to distinguish our model from the human. They will also provide a score (i.e. 0, 1, 2, the larger the better) indicating the quality of the answer. We propose strategies to monitor the quality of this evaluation process. The experiments show that in 64.7% of cases, the human judges cannot distinguish our model from humans. The average score is 1.454 (1.918 for human). The details of this work, including the FM-IQA dataset, can be found on the project page: \url{http://idl.baidu.com/FM-IQA.html}.

****

On the Pseudo-Dimension of Nearly Optimal Auctions

Jamie H. Morgenstern, Tim Roughgarden

This paper develops a general approach, rooted in statistical learning theory, to learning an approximately revenue-maximizing auction from data. We introduce t-level auctions to interpolate between simple auctions, such as welfare maximization with reserve prices, and optimal auctions, thereby balancing the competing demands of expressivity and simplicity. We prove that such auctions have small representation error, in the sense that for every product distribution F over bidders’ valuations, there exists a t-level auction with small t and expected revenue close to optimal. We show that the set of t-level auctions has modest pseudo-dimension (for polynomial t) and therefore leads to small learning error. One consequence of our results is that, in arbitrary single-parameter settings, one can learn a mechanism with expected revenue arbitrarily close to optimal from a polynomial number of samples.

****

Fast Second Order Stochastic Backpropagation for Variational Inference

Kai Fan, Ziteng Wang, Jeff Beck, James Kwok, Katherine A. Heller

We propose a second-order (Hessian or Hessian-free) based optimization method for variational inference inspired by Gaussian backpropagation, and argue that quasi-Newton optimization can be developed as well. This is accomplished by generalizing the gradient computation in stochastic backpropagation via a reparametrization trick with lower complexity. As an illustrative example, we apply this approach to the problems of Bayesian logistic regression and variational auto-encoder (VAE). Additionally, we compute bounds on the estimator variance of intractable expectations for the family of Lipschitz continuous function. Our method is practical, scalable and model free. We demonstrate our method on several real-world datasets and provide comparisons with other stochastic gradient methods to show substantial enhancement in convergence rates.

****

Cross-Domain Matching for Bag-of-Words Data via Kernel Embeddings of Latent Distributions

Yuya Yoshikawa, Tomoharu Iwata, Hiroshi Sawada, Takeshi Yamada

We propose a kernel-based method for finding matching between instances across different domains, such as multilingual documents and images with annotations. Each instance is assumed to be represented as a multiset of features, e.g., a bag-of-words representation for documents. The major difficulty in finding cross-domain relationships is that the similarity between instances in different domains cannot be directly measured. To overcome this difficulty, the proposed method embeds all the features of different domains in a shared latent space, and regards each instance as a distribution of its own features in the shared latent space. To represent the distributions efficiently and nonparametrically, we employ the framework of the kernel embeddings of distributions. The embedding is estimated so as to minimize the difference between distributions of paired instances while keeping unpaired instances apart. In our experiments, we show that the proposed method can achieve high performance on finding correspondence between multi-lingual Wikipedia articles, between documents and tags, and between images and tag

s.
****

Learning Large-Scale Poisson DAG Models based on OverDispersion Scoring
Gunwoong Park, Garvesh Raskutti
In this paper, we address the question of identifiability and learning algorithms for large-scale Poisson Directed Acyclic Graphical (DAG) models. We define general Poisson DAG models as models where each node is a Poisson random variable with rate parameter depending on the values of the parents in the underlying DAG. First, we prove that Poisson DAG models are identifiable from observational data, and present a polynomial-time algorithm that learns the Poisson DAG model under suitable regularity conditions. The main idea behind our algorithm is based on overdispersion, in that variables that are conditionally Poisson are overdispersed relative to variables that are marginally Poisson. Our algorithms exploits overdispersion along with methods for learning sparse Poisson undirected graphical models for faster computation. We provide both theoretical guarantees and simulation results for both small and large-scale DAGs.
****

Local Causal Discovery of Direct Causes and Effects
Tian Gao, Qiang Ji
We focus on the discovery and identification of direct causes and effects of a target variable in a causal network. State-of-the-art algorithms generally need to find the global causal structures in the form of complete partial directed acyclic graphs in order to identify the direct causes and effects of a target variable. While these algorithms are effective, it is often unnecessary and wasteful to find the global structures when we are only interested in one target variable (such as class labels). We propose a new local causal discovery algorithm, called Causal Markov Blanket (CMB), to identify the direct causes and effects of a target variable based on Markov Blanket Discovery. CMB is designed to conduct causal discovery among multiple variables, but focuses only on finding causal relationships between a specific target variable and other variables. Under standard assumptions, we show both theoretically and experimentally that the proposed local causal discovery algorithm can obtain the comparable identification accuracy as global methods but significantly improve their efficiency, often by more than one order of magnitude.
****

Recognizing retinal ganglion cells in the dark
Emile Richard, Georges A. Goetz, E.J. Chichilnisky
Many neural circuits are composed of numerous distinct cell types that perform different operations on their inputs, and send their outputs to distinct targets. Therefore, a key step in understanding neural systems is to reliably distinguish cell types. An important example is the retina, for which present-day techniques for identifying cell types are accurate, but very labor-intensive. Here, we develop automated classifiers for functional identification of retinal ganglion cells, the output neurons of the retina, based solely on recorded voltage patterns on a large scale array. We use per-cell classifiers based on features extracted from electrophysiological images (spatiotemporal voltage waveforms) and interspike intervals (autocorrelations). These classifiers achieve high performance in distinguishing between the major ganglion cell classes of the primate retina, but fail in achieving the same accuracy in predicting cell polarities (ON vs. OFF). We then show how to use indicators of functional coupling within populations of ganglion cells (cross-correlation) to infer cell polarities with a matrix completion algorithm. This can result in accurate, fully automated methods for cell type classification.
****